

# Maximizing the Conditional Expected Reward for Reaching the Goal

(extended version)<sup>★</sup>

Christel Baier, Joachim Klein, Sascha Klüppelholz, Sascha Wunderlich

Institute for Theoretical Computer Science  
Technische Universität Dresden, Germany

**Abstract.** The paper addresses the problem of computing maximal conditional expected accumulated rewards until reaching a target state (briefly called *maximal conditional expectations*) in finite-state Markov decision processes where the condition is given as a reachability constraint. Conditional expectations of this type can, e.g., stand for the maximal expected termination time of probabilistic programs with non-determinism, under the condition that the program eventually terminates, or for the worst-case expected penalty to be paid, assuming that at least three deadlines are missed. The main results of the paper are (i) a polynomial-time algorithm to check the finiteness of maximal conditional expectations, (ii) PSPACE-completeness for the threshold problem in acyclic Markov decision processes where the task is to check whether the maximal conditional expectation exceeds a given threshold, (iii) a pseudo-polynomial-time algorithm for the threshold problem in the general (cyclic) case, and (iv) an exponential-time algorithm for computing the maximal conditional expectation and an optimal scheduler.

## 1 Introduction

Stochastic shortest (or longest) path problems are a prominent class of optimization problems where the task is to find a policy for traversing a probabilistic graph structure such that the expected value of the generated paths satisfying a certain objective is minimal (or maximal). In the classical setting (see e.g. [14,33,23,28]), the underlying graph structure is given by a finite-state Markov decision process (MDP), i.e., a state-transition graph with nondeterministic choices between several actions for each of its non-terminal states, probability distributions specifying the probabilities for the successor states for each state-action pair and a reward function that assigns rational values to the state-action pairs. The stochastic shortest (longest) path problem asks to find a scheduler, i.e., a function that resolves the nondeterministic choices, possibly in a history-dependent way, which minimizes (maximizes) the expected accumulated reward until reaching a goal

---

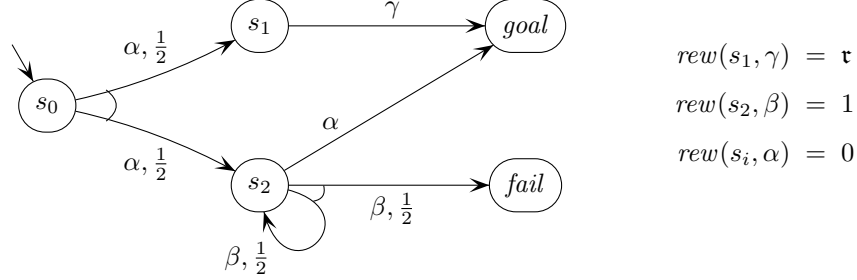
<sup>★</sup> The authors are supported by the DFG through the collaborative research centre HAEC (SFB 912), the Excellence Initiative by the German Federal and State Governments (cluster of excellence cfAED), the Research Training Group QuantLA (GRK 1763), and the DFG-project BA-1679/11-1.

state. To ensure the existence of the expectation for given schedulers, one often assumes that the given MDP is contracting, i.e., the goal is reached almost surely under all schedulers, in which case the optimal expected accumulated reward is achieved by a memoryless deterministic scheduler that optimizes the expectation from each state and is computable using a linear program with one variable per state (see e.g. [28]). The contraction assumption can be relaxed by requiring the existence of at least one scheduler that reaches the goal almost surely and taking the extremum over all those schedulers [14,23,15]. These algorithms and corresponding value or policy iteration approaches have been implemented in various tools and used in many application areas.

The restriction to schedulers that reach the goal almost surely, however, limits the applicability and significance of the results. First, the known algorithms for computing extremal expected accumulated rewards are not applicable for models where the probability for never visiting a goal state is positive under each scheduler. Second, statements about the expected rewards for schedulers that reach the goal with probability 1 are not sufficient to draw any conclusion for the best- or worst-case behavior, if there exist schedulers that miss the goal with positive probability. This motivates the consideration of *conditional stochastic path problems* where the task is to compute the optimal expected accumulated reward until reaching a goal state, under the condition that a goal state will indeed be reached and where the extrema are taken over all schedulers that reach the goal with positive probability. More precisely, we address here a slightly more general problem where we are given two sets  $F$  and  $G$  of states in an MDP  $\mathcal{M}$  with non-negative integer rewards and ask for the maximal expected accumulated reward until reaching  $F$ , under the condition that  $G$  will be visited (denoted  $\mathbb{E}_{\mathcal{M}, s_{init}}^{\max}(\Diamond F | \Diamond G)$  where  $s_{init}$  is the initial state of  $\mathcal{M}$ ). Computation schemes for conditional expectations of this type can, e.g., be used to answer the following questions (assuming the underlying model is a finite-state MDP):

- (Q1) What is the maximal termination time of a probabilistic and nondeterministic program, under the condition that the program indeed terminates?
- (Q2) What are the maximal expected costs of the repair mechanisms that are triggered in cases where a specific failure scenario occurs, under the condition that the failure scenario indeed occurs?
- (Q3) What is the maximal energy consumption, under the condition that all jobs of a given list will be successfully executed within one hour?

The relevance of question (Q1) and related problems becomes clear from the work [24,27,29,13,19] on the semantics of probabilistic programs where no guarantees for almost-sure termination can be given. Question (Q2) is natural for a worst-case analysis of resilient systems or other types of systems where conditional probabilities serve to provide performance guarantees on the protocols triggered in exceptional cases that appear with positive, but low probability. Question (Q3) is typical when the task is to study the trade-off between cost and utility functions (see e.g. [9]). Given the work on anonymity and related notions for information leakage using conditional probabilities in MDP-like models [7,20] or the formalization of posterior vulnerability as an expectation [4], the concept of



**Fig. 1.** MDP  $\mathcal{M}[\tau]$  for Example 1.1

conditional accumulated expected rewards might also be useful to specify the degree of protection of secret data or to study the trade-off between privacy and utility, e.g., using gain functions [5,3]. Other areas where conditional expectations play a crucial role are risk management where the conditional value-at-risk is used to formalize the expected loss under the assumption that very large losses occur [37,2] or regression analysis where conditional expectations serve to predict the relation between random variables [35].

*Example 1.1.* To illustrate the challenges for designing algorithms to compute maximal conditional expectations we regard the MDP  $\mathcal{M}[\tau]$  shown in Figure 1. The reward of the state-action pair  $(s_1, \gamma)$  is given by a reward parameter  $\tau \in \mathbb{N}$ . Let  $s_{init} = s_0$  be the initial state and  $F = G = \{\text{goal}\}$ . The only nondeterministic choice is in state  $s_2$ , while states  $s_0$  and  $s_1$  behave purely probabilistic and *goal* and *fail* are trap states. Given a scheduler  $\mathfrak{S}$ , we write  $\mathbb{CE}^{\mathfrak{S}}$  for the conditional expectation  $\mathbb{E}_{\mathcal{M}[\tau], s_0}^{\mathfrak{S}}(\Diamond \text{goal} | \Diamond \text{goal})$ . (See also Section 2 for our notations.) For the two memoryless schedulers that choose  $\alpha$  resp.  $\beta$  in state  $s_2$  we have:

$$\mathbb{CE}^{\alpha} = \frac{\frac{1}{2} \cdot \tau + \frac{1}{2} \cdot 0}{\frac{1}{2} + \frac{1}{2}} = \frac{\tau}{2} \quad \text{and} \quad \mathbb{CE}^{\beta} = \frac{\frac{1}{2} \cdot \tau + 0}{\frac{1}{2} + 0} = \tau$$

We now regard the schedulers  $\mathfrak{S}_n$  for  $n = 1, 2, \dots$  that choose  $\beta$  for the first  $n$  visits of  $s_2$  and action  $\alpha$  for the  $(n+1)$ -st visit of  $s_2$ . Then:

$$\mathbb{CE}^{\mathfrak{S}_n} = \frac{\frac{1}{2} \cdot \tau + \frac{1}{2} \cdot \frac{1}{2^n} \cdot n}{\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2^n}} = \tau + \frac{n - \tau}{2^n + 1}$$

Thus,  $\mathbb{CE}^{\mathfrak{S}_n} > \mathbb{CE}^{\beta}$  iff  $n > \tau$ , and the maximum is achieved for  $n = \tau + 2$ .

This example illustrates three phenomena that distinguish conditional and unconditional expected accumulated rewards and make reasoning about maximal conditional expectations harder than about unconditional ones. First, optimal schedulers for  $\mathcal{M}[\tau]$  need a counter for the number of visits in state  $s_2$ . Hence, memoryless schedulers are not powerful enough to maximize the conditional expectation. Second, while the maximal conditional expectation for  $\mathcal{M}[\tau]$  with initial state  $s_{init} = s_0$  is finite, the maximal conditional expectation for  $\mathcal{M}[\tau]$  with starting state  $s_2$  is infinite as:

$$\sup_{n \in \mathbb{N}} \mathbb{E}_{\mathcal{M}[\tau], s_2}^{\mathfrak{S}_n} (\Diamond goal | \Diamond goal) = \sup_{n \in \mathbb{N}} \frac{\frac{n}{2^n}}{\frac{1}{2^n}} = \infty$$

Third, as  $\mathfrak{S}_2$  maximizes the conditional expected accumulated reward for  $\tau = 0$ , while  $\mathfrak{S}_3$  is optimal for  $\tau = 1$ , optimal decisions for paths ending in state  $s_2$  depend on the reward value  $r$  of the  $\gamma$ -transition from state  $s_1$ , although state  $s_1$  is not reachable from  $s_2$ . Thus, optimal decisions for a path  $\pi$  do not only depend on the past (given by  $\pi$ ) and possible future (given by the sub-MDP that is reachable from  $\pi$ 's last state), but require global reasoning. ■

The main results of this paper are the following theorems. We write  $\mathbb{CE}^{\max}$  for the maximal conditional expectation, i.e., the supremum of the conditional expectations  $\mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond F | \Diamond G)$ , when ranging over all schedulers  $\mathfrak{S}$  where  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond G)$  is positive and  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond F | \Diamond G) = 1$ . (See also Section 2 for our notations.)

**Theorem 1 (Checking finiteness and upper bound)** *There is a polynomial-time algorithm that checks if  $\mathbb{CE}^{\max}$  is finite. If so, an upper bound  $\mathbb{CE}^{\text{ub}}$  for  $\mathbb{CE}^{\max}$  is computable in pseudo-polynomial time for the general case and in polynomial time if  $F = G$  and  $\Pr_{\mathcal{M}, s}^{\min} (\Diamond G) > 0$  for all states  $s$  with  $s \models \exists \Diamond G$ .*

The threshold problem asks whether the maximal conditional expectation exceeds or misses a given rational threshold  $\vartheta$ .

**Theorem 2 (Threshold problem)** *The problem “does  $\mathbb{CE}^{\max} \bowtie \vartheta$  hold?” (where  $\bowtie \in \{>, \geq, <, \leq\}$ ) is PSPACE-hard and solvable in exponential (even pseudo-polynomial) time. It is PSPACE-complete for acyclic MDPs.*

For the computation of an optimal scheduler, we suggest an iterative scheduler-improvement algorithm that interleaves calls of the threshold algorithm with linear programming techniques to handle zero-reward actions. This yields:

**Theorem 3 (Computing optimal schedulers)** *The value  $\mathbb{CE}^{\max}$  and an optimal scheduler  $\mathfrak{S}$  are computable in exponential time.*

Algorithms for checking finiteness and computing an upper bound (Theorem 1) will be sketched in Sections 3. Section 4 presents a pseudo-polynomial threshold algorithm and a polynomially space-bounded algorithm for acyclic MDPs (Theorem 2) as well as an exponential-time computation scheme for the construction of an optimal scheduler (Theorem 3). Further details, soundness proofs and a proof for the PSPACE-hardness as stated in Theorem 2 can be found in the appendix. The general feasibility of the algorithms will be shown by experimental studies with a prototypical implementation (for details, see Appendix K).

**Related work.** Although conditional expectations appear rather naturally in many applications and despite the large amount of publications on variants of stochastic path problems and other forms of expectations in MDPs (see e.g. [17, 34]), we are not aware that they have been addressed in the context of MDPs. Computation schemes for extremal conditional probabilities  $\Pr^{\max}(\varphi|\psi)$  or  $\Pr^{\min}(\varphi|\psi)$  where both the objective  $\varphi$  and the assumption  $\psi$  are path properties

specified in some temporal logic have been studied in [8,6,11]. For reachability properties  $\varphi$  and  $\psi$ , the algorithm of [8,6] has exponential time complexity, while the algorithm of [11] runs in polynomial time. Although the approach of [11] is not applicable for calculating maximal conditional expectations (see Appendix B), it can be used to compute an upper bound for  $\mathbb{CE}^{\max}$  (see Section 3). Conditional expected rewards in Markov chains can be computed using the rescaling technique of [11] for finite Markov chains or the approximation techniques of [18,1] for certain classes of infinite-state Markov chains. The conditional weakest precondition operator of [29] yields a technique to compute conditional expected rewards for purely probabilistic programs (without non-determinism).

## 2 Preliminaries

We briefly summarize our notations used for Markov decision processes. Further details can be found in textbooks, see e.g. [33,28] or Chapter 10 in [10].

A *Markov decision process* (MDP) is a tuple  $\mathcal{M} = (S, Act, P, s_{init}, rew)$  where  $S$  is a finite set of states,  $Act$  a finite set of actions,  $s_{init} \in S$  the initial state,  $P : S \times Act \times S \rightarrow [0, 1] \cap \mathbb{Q}$  is the transition probability function and  $rew : S \times Act \rightarrow \mathbb{N}$  the reward function. We require that  $\sum_{s' \in S} P(s, \alpha, s') \in \{0, 1\}$  for all  $(s, \alpha) \in S \times Act$ . We write  $Act(s)$  for the set of actions that are enabled in  $s$ , i.e.,  $\alpha \in Act(s)$  iff  $P(s, \alpha, \cdot)$  is not the null function. State  $s$  is called a *trap* if  $Act(s) = \emptyset$ . The paths of  $\mathcal{M}$  are finite or infinite sequences  $s_0 \alpha_0 s_1 \alpha_1 s_2 \alpha_2 \dots$  where states and actions alternate such that  $P(s_i, \alpha_i, s_{i+1}) > 0$  for all  $i \geq 0$ . A path  $\pi$  is called *maximal* if it is either infinite or finite and its last state is a trap. If  $\pi = s_0 \alpha_0 s_1 \alpha_1 s_2 \alpha_2 \dots \alpha_{k-1} s_k$  is finite then  $rew(\pi) = rew(s_0, \alpha_0) + rew(s_1, \alpha_1) + \dots + rew(s_{k-1}, \alpha_{k-1})$  denotes the accumulated reward and  $first(\pi) = s_0$ ,  $last(\pi) = s_k$  its first resp. last state. The *size* of  $\mathcal{M}$ , denoted  $size(\mathcal{M})$ , is the sum of the number of states plus the total sum of the logarithmic lengths of the non-zero probability values  $P(s, \alpha, s')$  and the reward values  $rew(s, \alpha)$ .<sup>1</sup>

An *end component* of  $\mathcal{M}$  is a strongly connected sub-MDP. End components can be formalized as pairs  $\mathcal{E} = (E, \mathfrak{A})$  where  $E$  is a nonempty subset of  $S$  and  $\mathfrak{A}$  a function that assigns to each state  $s \in E$  a nonempty subset of  $Act(s)$  such that the graph induced by  $\mathcal{E}$  is strongly connected.

A (*randomized*) *scheduler* for  $\mathcal{M}$ , often also called policy or adversary, is a function  $\mathfrak{S}$  that assigns to each finite path  $\pi$  where  $last(\pi)$  is not a trap a probability distribution over  $Act(last(\pi))$ .  $\mathfrak{S}$  is called *memoryless* if  $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$  for all finite paths  $\pi, \pi'$  with  $last(\pi) = last(\pi')$ , in which case  $\mathfrak{S}$  can be viewed as a function that assigns to each non-trap state  $s$  a distribution over  $Act(s)$ .  $\mathfrak{S}$  is called *deterministic* if  $\mathfrak{S}(\pi)$  is a Dirac distribution for each path  $\pi$ , in which case  $\mathfrak{S}$  can be viewed as a function that assigns an action to each finite path  $\pi$  where  $last(\pi)$  is not a trap. We write  $\Pr_{\mathcal{M}, s}^{\mathfrak{S}}$  or briefly  $\Pr_s^{\mathfrak{S}}$  to denote the

<sup>1</sup> The logarithmic length of an integer  $n$  is the number of bits required for a representation of  $n$  as a binary number. The logarithmic length of a rational number  $a/b$  is defined as the sum of the logarithmic lengths of its numerator  $a$  and its denominator  $b$ , assuming that  $a$  and  $b$  are coprime integers and  $b$  is positive.

probability measure induced by  $\mathfrak{S}$  and  $s$ . Given a measurable set  $\psi$  of maximal paths, then  $\Pr_{\mathcal{M},s}^{\min}(\psi) = \inf_{\mathfrak{S}} \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\psi)$  and  $\Pr_{\mathcal{M},s}^{\max}(\psi) = \sup_{\mathfrak{S}} \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\psi)$ . We will use LTL-like notations to specify measurable sets of maximal paths. For these it is well-known that optimal deterministic schedulers exists. If  $\psi$  is a reachability condition then even optimal deterministic memoryless schedulers exist.

Let  $\emptyset \neq F \subseteq S$ . For a comparison operator  $\bowtie \in \{=, >, \geq, <, \leq\}$  and  $r \in \mathbb{N}$ ,  $\Diamond^{\bowtie r} F$  denotes the event “reaching  $F$  along some finite path  $\pi$  with  $\text{rew}(\pi) \bowtie r$ ”. The notation  $\Diamond F$  will be used for the random variable that assigns to each maximal path  $\varsigma$  in  $\mathcal{M}$  the reward  $\text{rew}(\pi)$  of the shortest prefix  $\pi$  of  $\varsigma$  where  $\text{last}(\pi) \in F$ . If  $\varsigma \not\models \Diamond F$  then  $(\Diamond F)(\varsigma) = \infty$ . If  $s \in S$  then  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F)$  denotes the expectation of  $\Diamond F$  in  $\mathcal{M}$  with starting state  $s$  under  $\mathfrak{S}$ , which is infinite if  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F) < 1$ .  $\mathbb{E}_{\mathcal{M},s}^{\max}(\Diamond F) \in \mathbb{R} \cup \{\pm\infty\}$  stands for  $\sup_{\mathfrak{S}} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F)$  where the supremum is taken over all schedulers  $\mathfrak{S}$  with  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F) = 1$ . Let  $\psi$  be a measurable set of maximal paths.  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F|\psi)$  stands for the expectation of  $\Diamond F$  w.r.t. the conditional probability measure  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\cdot|\psi)$  given by  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\varphi|\psi) = \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\varphi \wedge \psi) / \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\psi)$ .  $\mathbb{E}_{\mathcal{M},s}^{\max}(\Diamond F|\psi)$  is the supremum of  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F|\psi)$  where  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\psi) > 0$  and  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F|\psi) = 1$ , and  $\Pr_{\mathcal{M},s}^{\max}(\varphi|\psi) = \sup_{\mathfrak{S}} \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\varphi|\psi)$  where  $\mathfrak{S}$  ranges over all schedulers with  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\psi) > 0$  and  $\sup \emptyset = -\infty$ .

For the remainder of this paper, we suppose that two nonempty subsets  $F$  and  $G$  of  $S$  are given such that  $\Pr_{\mathcal{M},s}^{\max}(\Diamond F|\Diamond G) = 1$ . The task addressed in this paper is to compute the maximal conditional expectation given by:

$$\mathbb{CE}_{\mathcal{M},s}^{\max} \stackrel{\text{def}}{=} \sup_{\mathfrak{S}} \mathbb{CE}_{\mathcal{M},s}^{\mathfrak{S}} \in \mathbb{R} \cup \{\infty\} \quad \text{where} \quad \mathbb{CE}_{\mathcal{M},s}^{\mathfrak{S}} = \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F|\Diamond G)$$

Here,  $\mathfrak{S}$  ranges over all schedulers  $\mathfrak{S}$  with  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond G) > 0$  and  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F|\Diamond G) = 1$ . If  $\mathcal{M}$  and its initial state are clear from the context, we often simply write  $\mathbb{CE}^{\max}$  resp.  $\mathbb{CE}^{\mathfrak{S}}$ . We assume that all states in  $\mathcal{M}$  are reachable from  $s_{\text{init}}$  and  $s_{\text{init}} \notin F \cup G$  (as  $\mathbb{CE}^{\max} = 0$  if  $s \in F$  and  $\mathbb{CE}^{\max} = \mathbb{E}_{\mathcal{M},s_{\text{init}}}^{\max}(\Diamond F)$  if  $s \in G \setminus F$ ).

### 3 Finiteness and upper bound

**Checking finiteness.** We sketch a polynomially time-bounded algorithm that takes as input an MDP  $\mathcal{M} = (S, \text{Act}, P, s_{\text{init}}, \text{rew})$  with two distinguished subsets  $F$  and  $G$  of  $S$  such that  $\Pr_{\mathcal{M},s_{\text{init}}}^{\max}(\Diamond F|\Diamond G) = 1$ . If  $\mathbb{CE}^{\max} = \mathbb{E}_{\mathcal{M},s_{\text{init}}}^{\max}(\Diamond F|\Diamond G) = \infty$  then the output is “no”. Otherwise, the output is an MDP  $\hat{\mathcal{M}} = (\hat{S}, \hat{\text{Act}}, \hat{P}, \hat{s}_{\text{init}}, \hat{\text{rew}})$  with two trap states *goal* and *fail* such that:

- (1)  $\mathbb{E}_{\mathcal{M},s_{\text{init}}}^{\max}(\Diamond F|\Diamond G) = \mathbb{E}_{\hat{\mathcal{M}},\hat{s}_{\text{init}}}^{\max}(\Diamond \text{goal}|\Diamond \text{goal})$ ,
- (2)  $\hat{s} \models \exists \Diamond \text{goal}$  and  $\Pr_{\hat{\mathcal{M}},\hat{s}}^{\min}(\Diamond(\text{goal} \vee \text{fail})) = 1$  for all states  $\hat{s} \in \hat{S} \setminus \{\text{fail}\}$ , and
- (3)  $\hat{\mathcal{M}}$  does not have critical schedulers where a scheduler  $\mathfrak{U}$  for  $\hat{\mathcal{M}}$  is said to be critical iff  $\Pr_{\hat{\mathcal{M}},\hat{s}_{\text{init}}}^{\mathfrak{U}}(\Diamond \text{fail}) = 1$  and there is a reachable positive  $\mathfrak{U}$ -cycle.<sup>2</sup>

<sup>2</sup> The latter means a  $\mathfrak{U}$ -path  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k$  where  $s_0 = \hat{s}_{\text{init}}$  and  $s_i = s_k$  for some  $i \in \{0, 1, \dots, k-1\}$  such that  $\text{rew}(s_j, \alpha_j) > 0$  for some  $j \in \{i, \dots, k-1\}$ .

We provide here the main ideas of the algorithms and refer to Appendix C for the details. The algorithm first transforms  $\mathcal{M}$  into an MDP  $\tilde{\mathcal{M}}$  that permits to assume  $F = G = \{goal\}$ . Intuitively,  $\tilde{\mathcal{M}}$  simulates  $\mathcal{M}$ , while operating in four modes: “normal mode”, “after  $G$ ”, “after  $F$ ” and “goal”.  $\tilde{\mathcal{M}}$  starts in normal mode where it behaves as  $\mathcal{M}$  as long as neither  $F$  nor  $G$  have been visited. If a  $G \setminus F$ -state has been reached in normal mode then  $\tilde{\mathcal{M}}$  switches to the mode “after  $G$ ”. Likewise, as soon as an  $F \setminus G$ -state has been reached in normal mode then  $\tilde{\mathcal{M}}$  switches to the mode “after  $F$ ”.  $\tilde{\mathcal{M}}$  enters the goal mode (consisting of a single trap state *goal*) as soon as a path fragment containing a state in  $F$  and a state in  $G$  has been generated. This is the case if  $\mathcal{M}$  visits an  $F$ -state in mode “after  $G$ ” or a  $G$ -state in mode “after  $F$ ”, or a state in  $F \cap G$  in the normal mode. The rewards in the normal mode and in mode “after  $G$ ” are precisely as in  $\mathcal{M}$ , while the rewards are 0 in all other cases. We then remove all states  $\tilde{s}$  in the “after  $G$ ” mode with  $\Pr_{\tilde{\mathcal{M}}, \tilde{s}}^{\max}(\Diamond goal) < 1$ , collapse all states  $\tilde{s}$  in  $\tilde{\mathcal{M}}$  with  $\tilde{s} \not\models \exists \Diamond goal$  into a single trap state called *fail* and add zero-reward transitions to *fail* from all states  $\tilde{s}$  that are not in the “after  $G$ ” mode and  $\Pr_{\tilde{\mathcal{M}}, \tilde{s}}^{\max}(\Diamond goal) = 0$ . Using techniques as in the unconditional case [23] we can check whether  $\tilde{\mathcal{M}}$  has positive end components, i.e., end components with at least one state-action pair  $(s, \alpha)$  with  $rew(s, \alpha) > 0$ . If so, then  $\mathbb{E}_{\tilde{\mathcal{M}}, s_{init}}^{\max}(\Diamond F | \Diamond G) = \infty$ . Otherwise, we collapse each maximal end component of  $\tilde{\mathcal{M}}$  into a single state.

Let  $\hat{\mathcal{M}}$  denote the resulting MDP. It satisfies (1) and (2). Property (3) holds iff  $\mathbb{E}_{\hat{\mathcal{M}}, s_{init}}^{\max}(\Diamond goal | \Diamond goal) < \infty$ . This condition can be checked in polynomial time using a graph analysis in the sub-MDP of  $\hat{\mathcal{M}}$  consisting of the states  $\hat{s}$  with  $\Pr_{\hat{\mathcal{M}}, \hat{s}}^{\min}(\Diamond goal) = 0$  (see Prop. C.8 and Appendix C.3).

**Computing an upper bound.** Due to the transformation used for checking finiteness of the maximal conditional expectation, we can now suppose that  $\mathcal{M} = \hat{\mathcal{M}}$ ,  $F = G = \{goal\}$  and that (2) and (3) hold. We now present a technique to compute an upper bound  $\mathbb{CE}^{\text{ub}}$  for  $\mathbb{CE}^{\max}$ . The upper bound will be used later to determine a saturation point from which on optimal schedulers behave memoryless (see Section 4).

We consider the MDP  $\mathcal{M}'$  simulating  $\mathcal{M}$ , while operating in two modes. In its first mode,  $\mathcal{M}'$  attaches the reward accumulated so far to the states. More precisely, the states of  $\mathcal{M}'$  in its first mode have the form  $\langle s, r \rangle \in S \times \mathbb{N}$  where  $0 \leq r \leq R$  and  $R = \sum_{s \in S'} \max\{rew_{\mathcal{M}'}(s, \alpha) : \alpha \in Act_{\mathcal{M}'}(s)\}$ . The initial state of  $\mathcal{M}'$  is  $s'_{init} = \langle s_{init}, 0 \rangle$ . The reward for the state-action pairs  $(\langle s, r \rangle, \alpha)$  where  $r + rew(s, \alpha) \leq R$  is 0. If  $\mathcal{M}'$  fires an action  $\alpha$  in state  $\langle s, r \rangle$  where  $r' \stackrel{\text{def}}{=} r + rew(s, \alpha) > R$  then it switches to the second mode, while earning reward  $r'$ . In its second mode  $\mathcal{M}'$  behaves as  $\mathcal{M}$  without additional annotations of the states and earning the same rewards as  $\mathcal{M}$ . From the states  $\langle goal, r \rangle$ ,  $\mathcal{M}'$  moves to *goal* with probability 1 and reward  $r$ . There is a one-to-one correspondence between the schedulers for  $\mathcal{M}$  and  $\mathcal{M}'$  and the switch from  $\mathcal{M}$  to  $\mathcal{M}'$  does not affect the probabilities and the accumulated rewards until reaching *goal*.

Let  $\mathcal{N}$  denote the MDP resulting from  $\mathcal{M}'$  by adding reset-transitions from *fail* (as a state of the second mode) and the copies  $\langle fail, r \rangle$  in the first mode

to the initial state  $s'_{init}$ . The reward of all reset transitions is 0. The reset-mechanism has been taken from [11] where it has been introduced as a technique to compute maximal conditional probabilities for reachability properties. Intuitively,  $\mathcal{N}$  “discards” all paths of  $\mathcal{M}'$  that eventually enter *fail* and “redistributes” their probabilities to the paths that eventually enter the goal state. In this way,  $\mathcal{N}$  mimics the conditional probability measures  $\Pr_{\mathcal{M}', s'_{init}}^{\mathfrak{S}}(\cdot \mid \Diamond goal) = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\cdot \mid \Diamond goal)$  for prefix-independent path properties. Paths  $\pi$  from  $s_{init}$  to *goal* in  $\mathcal{M}$  are simulated in  $\mathcal{N}$  by paths of the form  $\varrho = \xi_1; \dots \xi_k; \pi$  where  $\xi_i$  is a cycle in  $\mathcal{N}$  with  $first(\xi_i) = s'_{init}$  and  $\xi_i$ 's last transition is a reset-transition from some fail-state to  $s'_{init}$ . Thus,  $rew(\pi) \leq rew_{\mathcal{N}}(\varrho)$ . The distinction between the first and second mode together with property (3) ensure that the new reset-transitions do not generate positive end components in  $\mathcal{N}$ . By the results of [23], the maximal unconditional expected accumulated reward in  $\mathcal{N}$  is finite and we have:

$$\mathbb{E}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal \mid \Diamond goal) = \mathbb{E}_{\mathcal{M}', s'_{init}}^{\max}(\Diamond goal \mid \Diamond goal) \leq \mathbb{E}_{\mathcal{N}, s'_{init}}^{\max}(\Diamond goal)$$

Hence, we can deal with  $\mathbb{CE}^{\text{ub}} = \mathbb{E}_{\mathcal{N}, s'_{init}}^{\max}(\Diamond goal)$ , which is computable in time polynomial in the size of  $\mathcal{N}$  by the algorithm proposed in [23]. As  $size(\mathcal{N}) = \Theta(R \cdot size(\mathcal{M}))$  we obtain a pseudo-polynomial time bound for the general case. If, however,  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond goal) > 0$  for all states  $s \in S \setminus \{fail\}$  then there is no need for the detour via  $\mathcal{M}'$  and we can apply the reset-transformation  $\mathcal{M} \rightsquigarrow \mathcal{N}$  by adding a reset-transition from *fail* to  $s_{init}$  with reward 0, in which case the upper bound  $\mathbb{CE}^{\text{ub}} = \mathbb{E}_{\mathcal{N}, s_{init}}^{\max}(\Diamond goal)$  is obtained in time polynomial in the size of  $\mathcal{M}$ . For details we refer to the proof of Prop. C.8 and Section C.4 in the appendix.

## 4 Threshold algorithm and computing optimal schedulers

In what follows, we suppose that  $\mathcal{M} = (S, Act, P, s_{init}, rew)$  is an MDP with two trap states *goal* and *fail* such that  $s \models \exists \Diamond goal$  for all states  $s \in S \setminus \{fail\}$  and  $\min_{s \in S} \Pr_{\mathcal{M}, s}^{\min}(\Diamond(goal \vee fail)) = 1$  and  $\mathbb{CE}^{\max} = \mathbb{E}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal \mid \Diamond goal) < \infty$ .

A scheduler  $\mathfrak{S}$  is said to be *reward-based* if  $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$  for all finite paths  $\pi, \pi'$  with  $(last(\pi), rew(\pi)) = (last(\pi'), rew(\pi'))$ . Thus, deterministic reward-based schedulers can be seen as functions  $\mathfrak{S} : S \times \mathbb{N} \rightarrow Act$ . Prop. D.1 in the appendix shows that  $\mathbb{CE}^{\max}$  equals the supremum of the values  $\mathbb{CE}^{\mathfrak{S}}$ , when ranging over all deterministic reward-based schedulers  $\mathfrak{S}$  with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$ .

The basis of our algorithms are the following two observations. First, there exists a saturation point  $\wp \in \mathbb{N}$  such that the optimal decision for all paths  $\pi$  with  $rew(\pi) \geq \wp$  is to maximize the probability for reaching the goal state (see Prop. 4.1 below). The second observation is a technical statement that will be used at several places. Let  $\rho, \theta, \zeta, r, x, y, z, p \in \mathbb{R}$  with  $0 \leq p, x, y, z \leq 1$ ,  $p > 0$ ,  $y > z$  and  $x + z > 0$  and let

$$A = \frac{\rho + p(ry + \theta)}{x + py}, \quad B = \frac{\rho + p(rz + \zeta)}{x + pz} \quad \text{and} \quad C = \max\{A, B\}$$

Then:

$$A \geq B \quad \text{iff} \quad r + \frac{\theta - \zeta}{y - z} \geq C \quad \text{iff} \quad \theta - (C - r)y \geq \zeta - (C - r)z \quad (\dagger)$$



and the analogous statement for  $>$  rather than  $\geq$ . This statement is a consequence of Lemma G.1 in the appendix. We will apply this observation in different nuances. To give an idea how to apply statement  $(\dagger)$ , suppose  $A = \mathbb{CE}^{\mathfrak{T}}$  and  $B = \mathbb{CE}^{\mathfrak{U}}$  where  $\mathfrak{T}$  and  $\mathfrak{U}$  are reward-based schedulers that agree for all paths  $\varrho$  that do not have a prefix  $\pi$  with  $\text{rew}(\pi) = r$  where  $\text{last}(\pi)$  is a non-trap state, in which case  $x$  denotes the probability for reaching *goal* from  $s_{\text{init}}$  along such a path  $\varrho$  and  $\rho$  stands for the corresponding partial expectation, while  $p$  denotes the probability of the paths  $\pi$  from  $s_{\text{init}}$  to some non-trap state with  $\text{rew}(\pi) = r$ . The crucial observation is that  $r + (\theta - \zeta)/(y - z)$  does not depend on  $x, \rho, p$ . Thus, if  $r + (\theta - \zeta)/(y - z) \geq \mathbb{CE}^{\text{ub}}$  for some upper bound  $\mathbb{CE}^{\text{ub}}$  of  $\mathbb{CE}^{\text{max}}$  then  $(\dagger)$  allows to conclude that  $\mathfrak{T}$ 's decisions for the state-reward pairs  $(s, r)$  are better than  $\mathfrak{U}$ , independent of  $x, \rho$  and  $p$ .

Let  $R \in \mathbb{N}$  and  $\mathfrak{S}, \mathfrak{T}$  be reward-based schedulers. The *residual* scheduler  $\mathfrak{S} \uparrow R$  is given by  $(\mathfrak{S} \uparrow R)(s, r) = \mathfrak{S}(s, R + r)$ .  $\mathfrak{S} \triangleleft_R \mathfrak{T}$  denotes the unique scheduler that agrees with  $\mathfrak{S}$  for all state-reward pairs  $(s, r)$  where  $r < R$  and  $(\mathfrak{S} \triangleleft_R \mathfrak{T}) \uparrow R = \mathfrak{T}$ . We write  $E_{\mathcal{M},s}^{\mathfrak{S}}$  for the *partial expectation*

$$E_{\mathcal{M},s}^{\mathfrak{S}} = \sum_{r=0}^{\infty} \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond^{-r} \text{goal}) \cdot r$$

Thus,  $E_{\mathcal{M},s}^{\mathfrak{T}} = E_{\mathcal{M},s}^{\mathfrak{T}}(\Diamond \text{goal})$  if  $\Pr_{\mathcal{M},s}^{\mathfrak{T}}(\Diamond \text{goal}) = 1$ , while  $E_{\mathcal{M},s}^{\mathfrak{T}} < \infty = E_{\mathcal{M},s}^{\mathfrak{T}}(\Diamond \text{goal})$  if  $\Pr_{\mathcal{M},s}^{\mathfrak{T}}(\Diamond \text{goal}) < 1$ .

**Proposition 4.1.** *There exists a natural number  $\wp$  (called saturation point of  $\mathcal{M}$ ) and a deterministic memoryless scheduler  $\mathfrak{M}$  such that:*

- (a)  $\mathbb{CE}^{\mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}$  for each scheduler  $\mathfrak{T}$  with  $\Pr_{\mathcal{M},s_{\text{init}}}^{\mathfrak{T}}(\Diamond \text{goal}) > 0$ , and
- (b)  $\mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\text{max}}$  for some deterministic reward-based scheduler  $\mathfrak{S}$  such that  $\Pr_{\mathcal{M},s_{\text{init}}}^{\mathfrak{S}}(\Diamond \text{goal}) > 0$  and  $\mathfrak{S} \uparrow \wp = \mathfrak{M}$ .

The proof of Prop. 4.1 (see Appendices E and F) is constructive and yields a polynomial-time algorithm for generating a scheduler  $\mathfrak{M}$  as in Prop. 4.1 and a pseudo-polynomial algorithm for the computation of a saturation point  $\wp$ .

Scheduler  $\mathfrak{M}$  maximizes the probability to reach *goal* from each state. If there are two or more such schedulers, then  $\mathfrak{M}$  is one where the conditional expected accumulated reward until reaching goal is maximal under all schedulers  $\mathfrak{U}$  with  $\Pr_{\mathcal{M},s}^{\mathfrak{U}}(\Diamond \text{goal}) = \Pr_{\mathcal{M},s}^{\text{max}}(\Diamond \text{goal})$  for all states  $s$ . Such a scheduler  $\mathfrak{M}$  is computable in polynomial time using linear programming techniques. (See Lemma E.14 in the appendix.)

The idea for the computation of the saturation point is to compute the threshold  $\wp$  above which the scheduler  $\mathfrak{M}$  becomes optimal. For this we rely on statement  $(\dagger)$  where  $\theta/y$  stands for the conditional expectation under  $\mathfrak{M}$ ,  $\zeta/z$  for the conditional expectation under an arbitrary scheduler  $\mathfrak{S}$  and  $C = \mathbb{CE}^{\text{ub}}$  is an upper bound of  $\mathbb{CE}^{\text{max}}$  (see Theorem 1), while  $r = \wp$  is the wanted value. More precisely, for  $s \in S$ , let  $\theta_s = E_{\mathcal{M},s}^{\mathfrak{M}}$ ,  $y_s = \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond \text{goal}) = \Pr_{\mathcal{M},s}^{\text{max}}(\Diamond \text{goal})$ . To compute a saturation point we determine the smallest value  $\wp \in \mathbb{N}$  such that

$$\theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s = \max_{\mathfrak{S}} (E_{\mathcal{M},s}^{\mathfrak{S}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond \text{goal}))$$

for all states  $s$  where  $\mathfrak{S}$  ranges over all schedulers for  $\mathcal{M}$ . In Appendix F we show that instead of the maximum over all schedulers  $\mathfrak{S}$  it suffices to take the local maximum over all “one-step-variants” of  $\mathfrak{M}$ . That is, a saturation point is obtained by  $\wp = \max\{\lceil \mathbb{CE}^{\text{ub}} - D \rceil, 0\}$  where

$$D = \min \{ (\theta_s - \theta_{s,\alpha}) / (y_s - y_{s,\alpha}) : s \in S, \alpha \in \text{Act}(s), y_{s,\alpha} < y_s \}$$

and  $y_{s,\alpha} = \sum_{t \in S} P(s, \alpha, t) \cdot y_t$  and  $\theta_{s,\alpha} = \text{rew}(s, \alpha) \cdot y_{s,\alpha} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_t$ .

*Example 4.2.* The so obtained saturation point for the MDP  $\mathcal{M}[\mathfrak{r}]$  in Figure 1 is  $\wp = \lceil \mathbb{CE}^{\text{ub}} + 1 \rceil$ . Note that only state  $s = s_2$  behaves nondeterministically, and  $\mathfrak{M}(s) = \alpha$ ,  $y_s = y_{s,\alpha} = 1$ ,  $\theta_s = \theta_{s,\alpha} = 0$ , while  $y_{s,\beta} = \theta_{s,\beta} = \frac{1}{2}$ . This yields  $D = (0 - \frac{1}{2}) / (1 - \frac{1}{2}) = -1$ . Thus,  $\wp \geq \mathfrak{r} + 2$  as  $\mathbb{CE}^{\text{ub}} \geq \mathbb{CE}^{\text{max}} > \mathfrak{r}$ . ■

The logarithmic length of  $\wp$  is polynomial in the size of  $\mathcal{M}$ . Thus, the value (i.e., the length of an unary encoding) of  $\wp$  can be exponential in  $\text{size}(\mathcal{M})$ . This is unavoidable as there are families  $(\mathcal{M}_k)_{k \in \mathbb{N}}$  of MDPs where the size of  $\mathcal{M}_k$  is in  $\mathcal{O}(k)$ , while  $2^k$  is a lower bound for the smallest saturation point of  $\mathcal{M}_k$ . This, for instance, applies to the MDPs  $\mathcal{M}_k = \mathcal{M}[2^k]$  where  $\mathcal{M}[\mathfrak{r}]$  is as in Figure 1. Recall from Example 1.1 that the scheduler  $\mathfrak{S}_{\mathfrak{r}+2}$  that selects  $\beta$  by the first  $\mathfrak{r}+2$  visits of  $s$  and  $\alpha$  for the  $(\mathfrak{r}+3)$ -rd visit of  $s$  is optimal for  $\mathcal{M}[\mathfrak{r}]$ . Hence, the smallest saturation point for  $\mathcal{M}[2^k]$  is  $2^k + 2$ .

**Threshold algorithm.** The input of the threshold algorithm is an MDP  $\mathcal{M}$  as above and a non-negative rational number  $\vartheta$ . The task is to generate a deterministic reward-based scheduler  $\mathfrak{S}$  with  $\mathfrak{S} \uparrow \wp = \mathfrak{M}$  (where  $\mathfrak{M}$  and  $\wp$  are as in Prop. 4.1) such that  $\mathbb{CE}^{\mathfrak{S}} > \vartheta$  if  $\mathbb{CE}^{\text{max}} > \vartheta$ , and  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$  if  $\mathbb{CE}^{\text{max}} = \vartheta$ . If  $\mathbb{CE}^{\text{max}} < \vartheta$  then the output of the threshold algorithm is “no”.<sup>3</sup>

The algorithm operates level-wise and determines *feasible* actions  $\text{action}(s, r)$  for all non-trap states  $s$  and  $r = \wp - 1, \wp - 2, \dots, 0$ , using the decisions  $\text{action}(\cdot, i)$  for the levels  $i \in \{r+1, \dots, \wp\}$  that have been treated before and linear programming techniques to treat zero-reward loops. In this context, feasibility is understood with respect to the following condition: If  $\mathbb{CE}^{\text{max}} \geq \vartheta$  where  $\geq \in \{>, \geq\}$  then there exists a reward-based scheduler  $\mathfrak{S}$  with  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$  and  $\mathfrak{S}(s, R) = \text{action}(s, \min\{\wp, R\})$  for all  $R \geq r$ .

The algorithm stores for each state-reward pair  $(s, r)$  the probabilities  $y_{s,r}$  to reach *goal* from  $s$  and the corresponding partial expectation  $\theta_{s,r}$  for the scheduler given by the decisions in the action table. The values for  $r = \wp$  are given by  $\text{action}(s, \wp) = \mathfrak{M}(s)$ ,  $y_{s,\wp} = \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond \text{goal})$  and  $\theta_{s,\wp} = \mathbb{E}_{\mathcal{M},s}^{\mathfrak{M}}$ . The candidates for the decisions at level  $r < \wp$  are given by the deterministic memoryless schedulers  $\mathfrak{P}$  for  $\mathcal{M}$ . We write  $\mathfrak{P}_+$  for the reward-based scheduler given by  $\mathfrak{P}_+(s, 0) = \mathfrak{P}(s)$  and  $\mathfrak{P}_+(s, i) = \text{action}(s, \min\{\wp, r+i\})$  for  $i \geq 1$ . Let  $y_{s,r,\mathfrak{P}} = \Pr_{\mathcal{M},s}^{\mathfrak{P}_+}(\Diamond \text{goal})$  and  $\theta_{s,r,\mathfrak{P}} = \mathbb{E}_{\mathcal{M},s}^{\mathfrak{P}_+}$  be the corresponding partial expectation.

<sup>3</sup> The threshold algorithm solves all four variants of the threshold problem. E.g.,  $\mathbb{CE}^{\text{max}} \leq \vartheta$  iff  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$ , while  $\mathbb{CE}^{\text{max}} < \vartheta$  iff the threshold algorithm returns “no”.

To determine feasible actions for level  $r$ , the threshold algorithm makes use of a variant of (†) stating that if  $\theta - (\vartheta - r)y \geq \zeta - (\vartheta - r)z$  and  $B \supseteq \vartheta$  then  $A \supseteq \vartheta$ , where  $A$  and  $B$  are as in (†) and the requirement  $y > z$  is dropped. Thus, the aim of the threshold algorithm is to compute a deterministic memoryless scheduler  $\mathfrak{P}^*$  for  $\mathcal{M}$  such that the following condition (\*) holds:

$$\theta_{s,r,\mathfrak{P}^*} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}^*} = \max_{\mathfrak{P}} (\theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}}) \quad (*)$$

Such a scheduler  $\mathfrak{P}^*$  is computable in time polynomial in the size of  $\mathcal{M}$  (without the explicit consideration of all schedulers  $\mathfrak{P}$  and their extensions  $\mathfrak{P}_+$ ) using the following linear program with one variable  $x_s$  for each state. The objective is to minimize  $\sum_{s \in S} x_s$  subject to the following conditions:

(1) If  $s \in S \setminus \{goal, fail\}$  then for each action  $\alpha \in Act(s)$  with  $rew(s, \alpha) = 0$ :

$$x_s \geq \sum_{t \in S} P(s, \alpha, t) \cdot x_t$$

(2) If  $s \in S \setminus \{goal, fail\}$  then for each action  $\alpha \in Act(s)$  with  $rew(s, \alpha) > 0$ :

$$x_s \geq \sum_{t \in S} P(s, \alpha, t) \cdot (\theta_{t,R} + rew(s, \alpha) \cdot y_{t,R} - (\vartheta - r) \cdot y_{t,R})$$

where  $R = \min\{\varphi, r + rew(s, \alpha)\}$

(3) For the trap states:  $x_{goal} = r - \vartheta$  and  $x_{fail} = 0$

This linear program has a unique solution  $(x_s^*)_{s \in S}$ . Let  $Act^*(s)$  denote the set of actions  $\alpha \in Act(s)$  such that the following constraints (E1) and (E2) hold:

$$(E1) \quad \text{If } rew(s, \alpha) = 0 \text{ then: } x_s^* = \sum_{t \in S} P(s, \alpha, t) \cdot x_t^*$$

$$(E2) \quad \text{If } rew(s, \alpha) > 0 \text{ and } R = \min\{\varphi, r + rew(s, \alpha)\} \text{ then:}$$

$$x_s^* = \sum_{t \in S} P(s, \alpha, t) \cdot (\theta_{t,R} + rew(s, \alpha) \cdot y_{t,R} - (\vartheta - r) \cdot y_{t,R})$$

Let  $\mathcal{M}^* = \mathcal{M}_{r,\vartheta}^*$  denote the MDP with state space  $S$  induced by the state-action pairs  $(s, \alpha)$  with  $\alpha \in Act^*(s)$  where the positive-reward actions are redirected to the trap states. Formally, for  $s, t \in S$ ,  $\alpha \in Act^*(s)$  we let  $P_{\mathcal{M}^*}(s, \alpha, t) = P(s, \alpha, t)$  if  $rew(s, \alpha) = 0$  and  $P_{\mathcal{M}^*}(s, \alpha, goal) = \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,R}$  and  $P_{\mathcal{M}^*}(s, \alpha, fail) = 1 - P_{\mathcal{M}^*}(s, \alpha, goal)$  if  $rew(s, \alpha) > 0$  and  $R = \min\{\varphi, r + rew(s, \alpha)\}$ . The reward structure of  $\mathcal{M}^*$  is irrelevant for our purposes.

A scheduler  $\mathfrak{P}^*$  satisfying (\*) is obtained by computing a memoryless deterministic scheduler for  $\mathcal{M}^*$  with  $\Pr_{\mathcal{M}^*,s}^{\mathfrak{P}^*}(\Diamond goal) = \Pr_{\mathcal{M}^*,s}^{\max}(\Diamond goal)$  for all states  $s$ . This scheduler  $\mathfrak{P}^*$  indeed provides feasible decisions for level  $r$ , i.e., if  $\mathbb{CE}^{\max} \supseteq \vartheta$  where  $\supseteq \in \{>, \geq\}$  then there exists a reward-based scheduler  $\mathfrak{S}$  with  $\mathbb{CE}^{\mathfrak{S}} \supseteq \vartheta$ ,  $\mathfrak{S}(s, r) = \mathfrak{P}^*(s)$  and  $\mathfrak{S}(s, R) = action(s, \min\{\varphi, R\})$  for all  $R > r$ .

The threshold algorithm then puts  $action(s, r) = \mathfrak{P}^*(s)$  and computes the values  $y_{s,r}$  and  $\theta_{s,r}$  as follows. Let  $T$  denote the set of states  $s \in S \setminus \{goal, fail\}$  where  $rew(s, \mathfrak{P}^*(s)) > 0$ . For  $s \in T$ , the values  $y_{s,r} = y_{s,r,\mathfrak{P}^*}$  and  $\theta_{s,r} = \theta_{s,r,\mathfrak{P}^*}$

can be derived directly from the results obtained for the previously treated levels  $r+1, \dots, \wp$  as we have:

$$y_{s,r} = \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,R} \quad \text{and} \quad \theta_{s,r} = \text{rew}(s, \alpha) \cdot y_{s,r} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_{t,R}$$

where  $\alpha = \mathfrak{P}^*(s)$  and  $R = \min\{\wp, r + \text{rew}(s, \alpha)\}$ . For the states  $s \in S \setminus T$ :

$$y_{s,r} = \sum_{t \in T} \Pr_{\mathcal{M},s}^{\mathfrak{P}^*}(\neg T \cup t) \cdot y_{t,r} \quad \text{and} \quad \theta_{s,r} = \sum_{t \in T} \Pr_{\mathcal{M},s}^{\mathfrak{P}^*}(\neg T \cup t) \cdot \theta_{t,r}$$

Having treated the last level  $r = 0$ , the output of the algorithm is as follows. Let  $\mathfrak{S}$  be the scheduler given by the action table  $\text{action}(\cdot)$ . For the conditional expectation we have  $\mathbb{CE}^{\mathfrak{S}} = \theta_{s_{\text{init}},0}/y_{s_{\text{init}},0}$  if  $y_{s_{\text{init}},0} > 0$ . If  $y_{s_{\text{init}},0} = 0$  or  $\theta_{s_{\text{init}},0}/y_{s_{\text{init}},0} < \wp$  then the algorithm returns the answer “no”. Otherwise, the algorithm returns  $\mathfrak{S}$ , in which case  $\mathbb{CE}^{\mathfrak{S}} > \wp$  or  $\mathbb{CE}^{\mathfrak{S}} = \wp = \mathbb{CE}^{\max}$ . Proofs for the soundness and the pseudo-polynomial time complexity are provided in Appendix G.

*Example 4.3.* For the MDP  $\mathcal{M}[\mathfrak{r}]$  is Example 1.1, scheduler  $\mathfrak{M}$  selects action  $\alpha$  for state  $s = s_2$ . Thus,  $\text{action}(s, \wp) = \alpha$  for the computed saturation point  $\wp \geq \mathfrak{r}+2$  (see Example 4.2). The threshold algorithm for each positive rational threshold  $\wp$  computes for each level  $r = \wp-1, \wp-2, \dots, 1, 0$  where  $\text{action}(s, r+1) = \alpha$ , the value  $x_s^* = \max\{r-\wp, \frac{1}{2} + \frac{1}{2}(r-\wp)\}$  and the action set  $\text{Act}^*(s) = \{\alpha\}$  if  $r > \wp+1$ ,  $\text{Act}^*(s) = \{\alpha, \beta\}$  if  $r = \wp+1$  and  $\text{Act}^*(s) = \{\beta\}$  if  $r < \wp+1$ . Thus, if  $n = \min\{\wp, \lceil \wp+1 \rceil\}$  then  $\text{action}(s, r) = \alpha$ ,  $y_{s,r} = 1$ ,  $\theta_{s,r} = 0$  for  $r \in \{n, \dots, \wp\}$ , while  $\text{action}(s, n-k) = \beta$ ,  $y_{s,n-k} = 1/2^k$ ,  $\theta_{s,n-k} = k/2^k$  for  $k = 1, \dots, n$ . That is, the threshold algorithm computes the scheduler  $\mathfrak{S}_n$  that selects  $\beta$  for the first  $n$  visits of  $s$  and  $\alpha$  for the  $(n+1)$ -st visit of  $s$ . Thus, if  $\mathfrak{r} \leq \wp < \mathfrak{r}+1$  then  $n = \mathfrak{r}+2$ , in which case the computed scheduler  $\mathfrak{S}_n$  is optimal (see Example 1.1). The returned answer depends on whether  $\wp \leq \mathbb{CE}^{\max}$ . If, for instance,  $\wp = \frac{\mathfrak{r}}{2}$  and  $\mathfrak{r} > 0$  is even then the threshold algorithm returns the scheduler  $\mathfrak{S}_n$  where  $n = \frac{\mathfrak{r}}{2}+1$ , whose conditional expectation is  $\mathfrak{r} - (\frac{\mathfrak{r}}{2}-1)/(2^{\frac{\mathfrak{r}}{2}+1}+1) > \frac{\mathfrak{r}}{2} = \wp$ . ■

*MDPs without zero-reward cycles and acyclic MDPs.* If  $\mathcal{M}$  does not contain zero-reward cycles then there is no need for the linear program. Instead we can use a topological sorting of the states in the graph of the sub-MDP consisting of zero-reward actions and determine a scheduler  $\mathfrak{P}^*$  satisfying (\*) directly. For acyclic MDPs, there is even no need for a saturation point. We can explore  $\mathcal{M}$  using a recursive procedure and determine feasible decisions for each reachable state-reward pair  $(s, r)$  on the basis of (\*). This yields a polynomially space-bounded algorithm to decide whether  $\mathbb{CE}^{\max} \geq \wp$  in acyclic MDPs. (See Appendix I.)

**Construction of an optimal scheduler.** Let  $\text{ThresAlgo}[\wp]$  denote the scheduler that is generated by calling the threshold algorithm for the threshold value  $\wp$ . A simple approach is to apply the threshold algorithm iteratively:

```

let  $\mathfrak{S}$  be the scheduler  $\mathfrak{M}$  as in Proposition 4.1;
REPEAT  $\wp := \mathbb{CE}^{\mathfrak{S}}$ ;  $\mathfrak{S} := \text{ThresAlgo}[\wp]$  UNTIL  $\wp = \mathbb{CE}^{\mathfrak{S}}$ ;
return  $\wp$  and  $\mathfrak{S}$ 

```

The above algorithm generates a sequence of deterministic reward-based schedulers that are memoryless from  $\wp$  on with strictly increasing conditional expectations. The number of such schedulers is bounded by  $md^\wp$  where  $md$  denotes the number of memoryless deterministic schedulers for  $\mathcal{M}$ . Hence, the algorithm terminates and correctly returns  $\mathbb{CE}^{\max}$  and an optimal scheduler. As  $md$  can be exponential in the number of states, this simple algorithm has double-exponential time complexity.

To obtain a (single) exponential-time algorithm, we seek for better (larger, but still promising) threshold values than the conditional expectation of the current scheduler. We propose an algorithm that operates level-wise and freezes optimal decisions for levels  $r = \wp, \wp-1, \wp-2, \dots, 1, 0$ . The algorithm maintains and successively improves a left-closed and right-open interval  $I = [A, B[$  with  $\mathbb{CE}^{\max} \in I$  and  $\mathbb{CE}^\mathfrak{S} \in I$  for the current scheduler  $\mathfrak{S}$ .

*Initialization.* The algorithm starts with the scheduler  $\mathfrak{S} = \text{ThresAlgo}[\mathbb{CE}^\mathfrak{M}]$  where  $\mathfrak{M}$  is as above. If  $\mathbb{CE}^\mathfrak{S} = \mathbb{CE}^\mathfrak{M}$  then the algorithm immediately terminates. Suppose now that  $\mathbb{CE}^\mathfrak{S} > \mathbb{CE}^\mathfrak{M}$ . The initial interval is  $I = [A, B[$  where  $A = \mathbb{CE}^\mathfrak{S}$  and  $B = \mathbb{CE}^{\text{ub}} + 1$  where  $\mathbb{CE}^{\text{ub}}$  is as in Theorem 1.

*Level-wise scheduler improvement.* The algorithm successively determines optimal decisions for the levels  $r = \wp-1, \wp-2, \dots, 1, 0$ . The treatment of level  $r$  consists of a sequence of scheduler-improvement steps where at the same time the interval  $I$  is replaced with proper sub-intervals. The current scheduler  $\mathfrak{S}$  has been obtained by the last successful run of the threshold algorithm, i.e., it has the form  $\mathfrak{S} = \text{ThresAlgo}[\vartheta]$  where  $\mathbb{CE}^\mathfrak{S} > \vartheta$ . Besides the decisions of  $\mathfrak{S}$  (i.e., the actions  $\mathfrak{S}(s, R)$  for all state-reward pairs  $(s, R)$  where  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and  $R \in \{0, 1, \dots, \wp\}$ ), the algorithm also stores the values  $y_{s,R}$  and  $\theta_{s,R}$  that have been computed in the threshold algorithm.<sup>4</sup> For the current level  $r$ , the algorithm also computes for each state  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and each action  $\alpha \in \text{Act}(s)$  the values  $y_{s,r,\alpha} = \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,R}$  and  $\theta_{s,r,\alpha} = \text{rew}(s, \alpha) \cdot y_{s,r,\alpha} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_{t,R}$  where  $R = \min\{\wp, r + \text{rew}(s, \alpha)\}$ .

*Scheduler-improvement step.* Let  $r$  be the current level,  $I = [A, B[$  the current interval and  $\mathfrak{S}$  the current scheduler with  $\mathbb{CE}^{\max} \in I$ . At the beginning of the scheduler-improvement step we have  $\mathbb{CE}^\mathfrak{S} = A$ . Let

$$\begin{aligned} \mathcal{I}_{\mathfrak{S},r} &= \left\{ r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} : s \in S \setminus \{\text{goal}, \text{fail}\}, \alpha \in \text{Act}(s), y_{s,r} > y_{s,r,\alpha} \right\} \\ \mathcal{I}_{\mathfrak{S},r}^\uparrow &= \{ d \in \mathcal{I}_{\mathfrak{S},r} : d \geq \mathbb{CE}^\mathfrak{S} \} \quad \mathcal{I}_{\mathfrak{S},r}^B = \{ d \in \mathcal{I}_{\mathfrak{S},r} : d < B \} \end{aligned}$$

Intuitively, the values in  $d \in \mathcal{I}_{\mathfrak{S},r}^B$  are the “most promising” threshold values, as according to statement (†) these are the points where the decision of the current scheduler  $\mathfrak{S}$  for some state-reward pair  $(s, r)$  can be improved, provided that  $\mathbb{CE}^{\max} > d$ . (Note that the values in  $\mathcal{I}_{\mathfrak{S},r} \setminus \mathcal{I}_{\mathfrak{S},r}^B$  can be discarded as  $\mathbb{CE}^{\max} < B$ .)

<sup>4</sup> As the decisions of the already treated levels are optimal, the values  $y_{s,R}$  and  $\theta_{s,R}$  for  $R \in \{r+1, \dots, \wp\}$  can be reused in the calls of the threshold algorithms. That is, the calls of the threshold algorithm that are invoked in the scheduler-improvement steps at level  $r$  can skip levels  $\wp, \wp-1, \dots, r+1$  and only need to process levels  $r, r-1, \dots, 1, 0$ .

The algorithm proceeds as follows. If  $\mathcal{I}_{\mathfrak{S},r}^B = \emptyset$  then no further improvements at level  $r$  are possible as the function  $\mathfrak{P}^* = \mathfrak{S}(\cdot, r)$  satisfies  $(*)$  for the (still unknown) value  $\vartheta = \mathbb{CE}^{\max}$ . See Lemma H.7 in the appendix. In this case:

- If  $r = 0$  then the algorithm terminates with the answer  $\mathbb{CE}^{\max} = \mathbb{CE}^{\mathfrak{S}}$  and  $\mathfrak{S}$  as an optimal scheduler.
- If  $r > 0$  then the algorithm goes to the next level  $r-1$  and performs the scheduler-improvement step for  $\mathfrak{S}$  at level  $r-1$ .

Suppose now that  $\mathcal{I}_{\mathfrak{S},r}^B$  is nonempty. Let  $\mathcal{K} = \mathcal{I}_{\mathfrak{S},r}^{\uparrow} \cup \{\mathbb{CE}^{\mathfrak{S}}\}$ . The algorithm seeks for the largest value  $\vartheta' \in \mathcal{K} \cap I$  such that  $\mathbb{CE}^{\max} \geq \vartheta'$ . More precisely, it successively calls the threshold algorithm for the threshold value  $\vartheta' = \max(\mathcal{K} \cap I)$  and performs the following steps for the generated scheduler  $\mathfrak{S}' = \text{ThresAlgo}[\vartheta']$ :

- If the result of the threshold algorithm is “no” and  $\Pr_{\mathcal{M},s_{\text{init}}}^{\mathfrak{S}'}(\Diamond \text{goal})$  is positive (in which case  $\mathbb{CE}^{\mathfrak{S}'} \leq \mathbb{CE}^{\max} < \vartheta'$ ), then:
  - If  $\mathbb{CE}^{\mathfrak{S}'} \leq A$  then the algorithm refines  $I$  by putting  $B := \vartheta'$ .
  - If  $\mathbb{CE}^{\mathfrak{S}'} > A$  then the algorithm refines  $I$  by putting  $A := \mathbb{CE}^{\mathfrak{S}'}$ ,  $B := \vartheta'$  and adds  $\mathbb{CE}^{\mathfrak{S}'}$  to  $\mathcal{K}$  (Note that then  $\mathbb{CE}^{\mathfrak{S}'} \in \mathcal{K} \cap I$ , while  $\mathbb{CE}^{\mathfrak{S}} \in \mathcal{K} \setminus I$ ).
- Suppose now that  $\mathbb{CE}^{\mathfrak{S}'} \geq \vartheta'$ . The algorithm terminates if  $\mathbb{CE}^{\mathfrak{S}'} = \vartheta'$ , in which case  $\mathfrak{S}'$  is optimal. Otherwise, i.e., if  $\mathbb{CE}^{\mathfrak{S}'} > \vartheta'$ , then the algorithm aborts the loop by putting  $\mathcal{K} := \emptyset$ , refines the interval  $I$  by putting  $A := \mathbb{CE}^{\mathfrak{S}'}$ , updates the current scheduler by setting  $\mathfrak{S} := \mathfrak{S}'$  and performs the next scheduler-improvement step.

The soundness proof and complexity analysis can be found in Appendix H, where (among others) we show that the scheduler-improvement step for schedulers  $\mathfrak{S}$  with  $\mathbb{CE}^{\mathfrak{S}} < \mathbb{CE}^{\max}$  terminates with some scheduler  $\mathfrak{S}'$  such that  $\mathbb{CE}^{\mathfrak{S}} < \mathbb{CE}^{\mathfrak{S}'}$ . The total number of calls of the threshold algorithm is in  $\mathcal{O}(\wp \cdot md \cdot |S| \cdot |\text{Act}|)$ . This yields an exponential time bound as stated in Theorem 3.

*Example 4.4.* We regard again the MDP  $\mathcal{M}[\mathfrak{r}]$  of Example 1.1 where we suppose  $\mathfrak{r}$  is positive and even. The algorithm first computes  $\mathbb{CE}^{\text{ub}}$  (see Section 3), a saturation point  $\wp \geq \mathfrak{r}+2$  (see Example 4.2), the scheduler  $\mathfrak{M}$ , its conditional expectation  $\mathbb{CE}^{\mathfrak{M}} = \frac{\mathfrak{r}}{2}$  and the scheduler  $\mathfrak{S} = \text{ThresAlgo}[\frac{\mathfrak{r}}{2}]$ . The initial interval is  $I = [A, B[$  where  $A = \mathbb{CE}^{\mathfrak{S}} = \mathfrak{r} - (\frac{\mathfrak{r}}{2}-1)/(2^{\frac{\mathfrak{r}}{2}+1}+1)$  (see Example 4.3) and  $B = \mathbb{CE}^{\text{ub}}+1$ . The scheduler improvement step for  $\mathfrak{S}$  at levels  $r = \wp-1, \dots, \mathfrak{r}+1$  determines the set  $\mathcal{I}_{\mathfrak{S},r} = \{r-1\}$  and calls the threshold algorithm for  $\vartheta' = r-1$ . These calls are not successful for  $r = \wp-1, \dots, \mathfrak{r}+2$ . That is, the scheduler  $\mathfrak{S}$  remains unchanged and the upper bound  $B$  is successively improved to  $r-1$ . At level  $r = \mathfrak{r}+1$ , the threshold algorithm is called for  $\vartheta' = \mathfrak{r}$ , which yields the optimal scheduler  $\mathfrak{S}' = \text{ThresAlgo}[\vartheta']$  (see Example 4.3). ■

**Implementation and experiments.** We have implemented the algorithms presented in this paper as a prototypical extension of the model checker PRISM [30,32] and carried out initial experiments to demonstrate the general feasibility of our approach (see <https://www.tcs.inf.tu-dresden.de/ALGI/PUB/TACAS17/> and Appendix K for details).

## 5 Conclusion

Although the switch to conditional expectations appears rather natural to escape from the limitations of known solutions for unconditional extremal expected accumulated rewards, to the best of our knowledge computation schemes for conditional expected accumulated rewards have not been addressed before. Our results show that new techniques are needed to compute maximal conditional expectations, as optimal schedulers might need memory and local reasoning in terms of the past and possible future is not sufficient (Example 1.1). The key observations for our algorithms are the existence of a saturation point  $\wp$  for the reward that has been accumulated so far, from which on optimal schedulers can behave memoryless, and a linear correlation between optimal decisions for all state-reward pairs  $(s, r)$  of the same reward level  $r$  (see (\*) and the linear program used in the threshold algorithm). The difficulty to reason about conditional expectations is also reflected in the achieved complexity-theoretic results stating that all variants of the threshold problem lie between PSPACE and EXPTIME. While PSPACE-completeness has been established for acyclic MDPs (Appendix I), the precise complexity for cyclic MDPs is still open. In contrast, optimal schedulers for unconditional expected accumulated rewards as well as for conditional reachability probabilities are computable in polynomial time [23,11].

Using standard automata-based approaches, our method can easily be generalized to compute maximal conditional expected rewards for regular co-safety conditions (rather than reachability conditions  $\Diamond G$ ) and/or where the accumulation of rewards is “controlled” by a deterministic finite automaton as in the logics considered in [16,12] (rather than  $\Diamond F$ ). In this paper, we restricted to MDPs with non-negative integer rewards. Non-negative rational rewards can be treated by multiplying all reward values with their least common multiple (Appendix J.1). In the case of acyclic MDPs, our methods are even applicable if the MDP has negative and positive rational rewards (Appendix J.2). By swapping the sign of all rewards, this yields a technique to compute minimal conditional expectations in acyclic MDPs. We expect that minimal conditional expectations in cyclic MDPs with non-negative rewards can be computed using similar algorithms as we suggested for maximal conditional expectations. This as well as MDPs with negative and positive rewards will be addressed in future work.

## References

1. P. A. Abdulla, N. B. Henda, and R. Mayr. Decisive Markov chains. *Logical Methods in Computer Science*, 3(4), 2007.
2. C. Acerbi and D. Tasche. Expected shortfall: A natural coherent alternative to value at risk. *Economic Notes*, 31(2):379–388, 2002.
3. M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi. On the information leakage of differentially-private mechanisms. *Journal of Computer Security*, 23(4):427–469, 2015.

4. M. S. Alvim, K. Chatzikokolakis, A. McIver, C. Morgan, C. Palamidessi, and G. Smith. Axioms for information leakage. In *Proc. Computer Security Foundations Symposium (CSF)*, pages 77–92. IEEE Computer Society, 2016.
5. M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, and G. Smith. Measuring information leakage using generalized gain functions. In *Proc. Computer Security Foundations Symposium (CSF)*, pages 265–279. IEEE Computer Society, 2012.
6. M. E. Andrés. *Quantitative Analysis of Information Leakage in Probabilistic and Nondeterministic Systems*. PhD thesis, UB Nijmegen, 2011.
7. M. E. Andrés, C. Palamidessi, P. van Rossum, and A. Sokolova. Information hiding in probabilistic concurrent systems. *Theoretical Computer Science*, 412(28):3072–3089, 2011.
8. M. E. Andrés and P. van Rossum. Conditional probabilities over probabilistic and nondeterministic systems. In *Proc. Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, volume 4963 of *LNCS*, pages 157–172. Springer, 2008.
9. C. Baier, C. Dubslaff, J. Klein, S. Klüppelholz, and S. Wunderlich. Probabilistic model checking for energy-utility analysis. In *Horizons of the Mind. A Tribute to Prakash Panangaden*, volume 8464 of *LNCS*, pages 96–123. Springer, 2014.
10. C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT Press, 2008.
11. C. Baier, J. Klein, S. Klüppelholz, and S. Märcker. Computing conditional probabilities in Markovian models efficiently. In *Proc. Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, volume 8413 of *LNCS*, pages 515–530. Springer, 2014.
12. C. Baier, J. Klein, S. Klüppelholz, and S. Wunderlich. Weight monitoring with linear temporal logic: Complexity and decidability. In *Proc. Computer Science Logic/Logic In Computer Science (CSL-LICS)*, pages 11:1–11:10. ACM, 2014.
13. G. Barthe, T. Espitau, L. M. F. Fioriti, and J. Hsu. Synthesizing probabilistic invariants via Doob’s decomposition. In *Proc. Computer Aided Verification (CAV)*, volume 9779 of *LNCS*, pages 43–61. Springer, 2016.
14. D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, 1991.
15. D. P. Bertsekas and H. Yu. Stochastic path problems under weak conditions. Technical report, M.I.T. Cambridge, 2016. Report LIDS 2909.
16. U. Boker, K. Chatterjee, T. A. Henzinger, and O. Kupferman. Temporal specifications with accumulative values. In *Proc. Logic in Computer Science (LICS)*, pages 43–52. IEEE Computer Society, 2011.
17. T. Brázdil, V. Brozek, K. Chatterjee, V. Forejt, and A. Kucera. Two views on multiple mean-payoff objectives in Markov decision processes. *Logical Methods in Computer Science*, 10(1), 2014.
18. T. Brázdil and A. Kucera. Computing the expected accumulated reward and gain for a subclass of infinite Markov chains. In *Proc. Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, volume 3821 of *LNCS*, pages 372–383. Springer, 2005.
19. K. Chatterjee, H. Fu, and A. K. Goharshady. Termination analysis of probabilistic programs through Positivstellensatz’s. In *Proc. Computer Aided Verification (CAV)*, volume 9779 of *LNCS*, pages 3–22. Springer, 2016.
20. K. Chatzikokolakis, C. Palamidessi, and C. Braun. Compositional methods for information-hiding. *Mathematical Structures in Computer Science*, 26(6):908–932, 2016.



21. F. Ciesinski, C. Baier, M. Größer, and J. Klein. Reduction techniques for model checking Markov decision processes. In *Proc. Quantitative Evaluation of Systems (QEST)*, pages 45–54. IEEE Computer Society Press, 2008.
22. C. Courcoubetis, M. Y. Vardi, P. Wolper, and M. Yannakakis. Memory-efficient algorithms for the verification of temporal properties. *Formal Methods in System Design*, 1(2):275–288, 1992.
23. L. de Alfaro. Computing minimum and maximum reachability times in probabilistic systems. In *Proc. Concurrency Theory (CONCUR)*, volume 1664 of *LNCS*, pages 66–81, 1999.
24. F. Gretz, J. Katoen, and A. McIver. Operational versus weakest pre-expectation semantics for the probabilistic guarded command language. *Performance Evaluation*, 73:110–132, 2014.
25. C. Haase and S. Kiefer. The odds of staying on budget. In *Proc. International Colloquium on Automata, Languages, and Programming (ICALP), Part II*, volume 9135 of *LNCS*, pages 234–246. Springer, 2015.
26. S. Haddad and B. Monmege. Reachability in MDPs: Refining convergence of value iteration. In *Proc. Reachability Problems (RP)*, volume 8762 of *LNCS*, pages 125–137. Springer, 2014.
27. N. Jansen, B. L. Kaminski, J. Katoen, F. Olmedo, F. Gretz, and A. McIver. Conditioning in probabilistic programming. In *Proc. Mathematical Foundations of Programming Semantics (MFPS)*, volume 319 of *Electronic Notes Theoretical Computer Science*, pages 199–216, 2015.
28. L. Kallenberg. *Markov Decision Processes*. Lecture Notes. University of Leiden, 2011.
29. J. Katoen, F. Gretz, N. Jansen, B. L. Kaminski, and F. Olmedo. Understanding probabilistic programs. In *Correct System Design*, volume 9360 of *LNCS*, pages 15–32. Springer, 2015.
30. M. Kwiatkowska, G. Norman, and D. Parker. PRISM 4.0: Verification of probabilistic real-time systems. In *Proc. Computer Aided Verification (CAV)*, volume 6806 of *LNCS*, pages 585–591. Springer, 2011.
31. M. Z. Kwiatkowska, G. Norman, and D. Parker. The PRISM benchmark suite. In *Proc. Quantitative Evaluation of Systems (QEST)*, pages 203–204. IEEE Computer Society, 2012.
32. PRISM model checker. <http://www.prismmodelchecker.org/>.
33. M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994.
34. M. Randour, J. Raskin, and O. Sankur. Variations on the stochastic shortest path problem. In *Proc. Verification, Model Checking, and Abstract Interpretation (VMCAI)*, volume 8931 of *LNCS*, pages 1–18. Springer, 2015.
35. G. Seber and A. Lee. *Linear Regression Analysis*. Wiley Series in Probability and Statistics, 2003.
36. M. Ummels and C. Baier. Computing quantiles in Markov reward models. In *Proc. Foundations of Software Science and Computation Structures (FOSSACS)*, volume 7794 of *LNCS*, pages 353–368. Springer, 2013.
37. S. Uryasev. Conditional value-at-risk: optimization algorithms and applications. In *Proc. Computational Intelligence and Financial Engineering (CIFER)*, pages 49–57. IEEE, 2000.

## APPENDIX

### Outline of the appendix.

- Appendix A (Relevant notations for the appendix, page 18) explains the notations used in the appendix.
- Appendix B (Extremal conditional probabilities, page 21) sketches the reset mechanism of [11] for computing maximal conditional probabilities (used in Section 3 to compute an upper bound  $\mathbb{CE}^{\text{ub}}$ ). It also provides an example illustrating why the reset method fails for conditional expectations.
- Appendix C (Finiteness and upper bound, page 22) deals with determining finiteness and the computation of an upper bound (Theorem 1).
- Appendix D (Deterministic reward-based schedulers are sufficient, page 36) shows that deterministic reward-based schedulers are sufficient for maximizing the conditional expectation in our setting.
- Appendix E (Existence of a saturation point and optimal schedulers, page 38) provides the proof for Proposition 4.1.
- Appendix F (Computing a saturation point, page 51) shows the correctness of the computation of a saturation point as described in Section 4.
- Appendix G (Threshold algorithm, page 58) provides additional details for the threshold algorithm as well as the proof of the soundness proof and the pseudo-polynomial time bound as stated in Theorem 2.
- Appendix H (Computing an optimal scheduler and the maximal conditional expectation, page 73) provides additional details for the scheduler improvement algorithm of Section 4 and the proof of Theorem 3.
- Appendix I (PSPACE-completeness for acyclic MPDs, page 84) provides the PSPACE-completeness proof of the threshold problem in acyclic MDPs as stated in Theorem 2.
- Appendix J (Rational and negative rewards, page 97) provides details for rational and negative rewards as mentioned in the conclusion.
- Appendix K (Implementation and experiments, page 101) provides details on our prototypical implementation in PRISM and our experiments.

## A Relevant notations for the appendix

**Notations for Markov decision processes.** A *Markov decision process* (MDP) is a tuple  $\mathcal{M} = (S, \text{Act}, P, s_{\text{init}}, \text{rew})$  where  $S$  is a finite set of states,  $\text{Act}$  a finite set of actions,  $s_{\text{init}} \in S$  the initial state,  $P : S \times \text{Act} \times S \rightarrow [0, 1] \cap \mathbb{Q}$  is the transition probability function and  $\text{rew} : S \times \text{Act} \rightarrow \mathbb{N}$  the reward function. We require that  $P(s, \alpha, S) \in \{0, 1\}$  for all  $(s, \alpha) \in S \times \text{Act}$  where, for  $F \subseteq S$ ,  $P(s, \alpha, F) = \sum_{t \in F} P(s, \alpha, t)$ . We write  $\text{Act}(s)$  for the set of actions that are enabled in  $s$ , i.e.,  $\alpha \in \text{Act}(s)$  iff  $P(s, \alpha, \cdot)$  is not the null function. State  $s$  is called a *trap* if  $\text{Act}(s) = \emptyset$ .

The paths of  $\mathcal{M}$  are finite or infinite sequences  $s_0 \alpha_0 s_1 \alpha_1 s_2 \alpha_2 \dots$  where states and actions alternate such that  $P(s_i, \alpha_i, s_{i+1}) > 0$  for all  $i \geq 0$ . A path

$\pi$  is called *maximal* if it is either infinite or finite and its last state is a trap. If  $\pi = s_0 \alpha_0 s_1 \alpha_1 s_2 \alpha_2 \dots \alpha_{k-1} s_k$  is finite then

- $rew(\pi) = rew(s_0, \alpha_0) + rew(s_1, \alpha_1) + \dots + rew(s_{k-1}, \alpha_{k-1})$  denotes the accumulated reward,
- $|\pi| = k$  its length (number of transitions),
- $prob(\pi) = P(s_0, \alpha_0, s_1) \cdot P(s_1, \alpha_1, s_2) \cdot \dots \cdot P(s_{k-1}, \alpha_{k-1}, s_k)$  its probability,
- $first(\pi) = s_0$ ,  $last(\pi) = s_k$  its first resp. last state.

The notation  $\pi_1; \pi_2$  is used to denote the concatenation of paths  $\pi_1$  and  $\pi_2$  with  $last(\pi_1) = first(\pi_2)$ , where  $\pi_1; \pi_2 = \pi_2$  if  $|\pi_1| = 0$ .

The *size* of  $\mathcal{M}$ , denoted  $size(\mathcal{M})$ , is the sum of the number of states plus the total sum of the logarithmic lengths of the non-zero probability values  $P(s, \alpha, s')$  and the reward values  $rew(s, \alpha)$ .

**Scheduler.** A (*randomized*) scheduler for  $\mathcal{M}$ , often also called policy or adversary, is a function  $\mathfrak{S}$  that assigns to each finite path  $\pi$  where  $last(\pi)$  is not a trap a probability distribution over  $Act(last(\pi))$ .  $\mathfrak{S}$  is called memoryless if  $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$  for all finite paths  $\pi, \pi'$  with  $last(\pi) = last(\pi')$ , in which case  $\mathfrak{S}$  can be viewed as a function that assigns to each non-trap state  $s$  a distribution over  $Act(s)$ .  $\mathfrak{S}$  is called deterministic if  $\mathfrak{S}(\pi)$  is a Dirac distribution for each path  $\pi$ , in which case  $\mathfrak{S}$  can be viewed as a function that assigns an action to each finite path  $\pi$  where  $last(\pi)$  is not a trap. Given a scheduler  $\mathfrak{S}$ , a  $\mathfrak{S}$ -path is any path that might arise when the nondeterministic choices in  $\mathcal{M}$  are resolved using  $\mathfrak{S}$ . Thus,  $\varsigma = s_0 \alpha_0 s_1 \alpha_1 \dots$  is a  $\mathfrak{S}$ -path iff  $\varsigma$  is a path and  $\mathfrak{S}(s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{k-1} s_k)(\alpha_k) > 0$  for all  $k \geq 0$ . If  $\pi$  is a finite  $\mathfrak{S}$ -path then  $\mathfrak{S} \uparrow \pi$  denotes the residual scheduler “ $\mathfrak{S}$  after  $\pi$ ” given by:

$$(\mathfrak{S} \uparrow \pi)(\varrho) = \mathfrak{S}(\pi; \varrho) \quad \text{if } \varrho \text{ is a finite path with } last(\pi) = first(\varrho)$$

The behavior of  $\mathfrak{S} \uparrow \pi$  for paths not starting in  $last(\pi)$  is irrelevant.

Let  $\mathfrak{S}$  and  $\mathfrak{U}$  be schedulers. If  $\pi$  is a finite  $\mathfrak{S}$ -path  $\pi$ , then  $\mathfrak{S} \triangleleft_{\pi} \mathfrak{U}$  denotes the unique scheduler  $\mathfrak{T}$  with  $\mathfrak{T} \uparrow \pi = \mathfrak{U}$  that behaves as  $\mathfrak{S}$  for all paths  $\varrho$  where  $\pi$  is not a prefix of  $\varrho$ . That is:

$$(\mathfrak{S} \triangleleft_{\pi} \mathfrak{U})(\varrho) = \begin{cases} \mathfrak{S}(\varrho) & : \text{if } \pi \text{ is not a prefix of } \varrho \\ \mathfrak{U}(\varrho') & : \text{if } \varrho = \pi; \varrho' \end{cases}$$

where  $\pi; s = \pi$ . Hence,  $(\mathfrak{S} \triangleleft_{\pi} \mathfrak{U})(\pi) = \mathfrak{U}(last(\pi))$ . If  $R \in \mathbb{N}$  then  $\mathfrak{S} \triangleleft_R \mathfrak{U}$  denotes the scheduler given by  $(\mathfrak{S} \triangleleft_R \mathfrak{U})(\varrho) = \mathfrak{S}(\varrho)$  if  $rew(\varrho) < R$  and  $(\mathfrak{S} \triangleleft_R \mathfrak{U})(\pi; \varrho') = \mathfrak{U}(\varrho')$  if  $rew(\pi) \geq R$  and  $rew(\pi') < R$  for all proper prefixes  $\pi'$  of  $\pi$ .

Scheduler  $\mathfrak{S}$  is said to be *reward-based* if  $\mathfrak{S}(\pi) = \mathfrak{S}(\pi')$  for all finite paths  $\pi, \pi'$  with  $rew(\pi) = rew(\pi')$  and  $last(\pi) = last(\pi')$ . Thus, deterministic reward-based schedulers can be viewed as function that assign actions to state-reward pairs. As stated in the main paper, for reward-based schedulers we use the notation  $\mathfrak{S} \uparrow R$  to denote the reward-based scheduler given by  $(\mathfrak{S} \uparrow R)(s, r) = \mathfrak{S}(s, R+r)$ . We use the notation  $\mathfrak{S} \uparrow(s, R)$  to denote the scheduler  $\mathfrak{S} \uparrow \pi$  for each/some finite  $\mathfrak{S}$ -path

$\pi$  with  $\text{rew}(\pi) = R$  and  $\text{last}(\pi) = s$ .<sup>5</sup> Clearly, if  $\mathfrak{S}$  and  $\mathfrak{U}$  are reward-based schedulers and  $R \in \mathbb{N}$  then  $\mathfrak{S} \triangleleft_R \mathfrak{U}$  is a reward-based scheduler too. In this case,  $(\mathfrak{S} \triangleleft_R \mathfrak{U})(s, r) = \mathfrak{S}(s, r)$  for  $r < R$  and  $(\mathfrak{S} \triangleleft_R \mathfrak{U})(s, r) = \mathfrak{U}(s, r - R)$  for  $r \geq R$ . Hence,  $(\mathfrak{S} \triangleleft_R \mathfrak{U}) \uparrow R = \mathfrak{U}$ .

**Probability measure.** We write  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}$  or briefly  $\text{Pr}_s^{\mathfrak{S}}$  to denote the probability measure induced by  $\mathfrak{S}$  and  $s$ . The underlying sigma-algebra is the one that is generated by the cylinder sets of the finite paths starting in  $s$  where the cylinder set  $\text{Cyl}(\pi)$  of  $\pi$  consists of all maximal paths  $\varsigma$  that are extensions of  $\pi$ . Then,  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}$  is the unique probability measure such that for each finite path  $\pi = s_0 \alpha_0 s_1 \alpha_1 s_2 \alpha_2 \dots \alpha_{k-1} s_k$ :

$$\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\text{Cyl}(\pi)) = \text{prob}(\pi) \cdot \prod_{i=0}^{k-1} \mathfrak{S}(\text{pref}(\pi, i))(\alpha_i)$$

where  $\text{pref}(\pi, i) = s_0 \alpha_0 s_1 \alpha_1 s_2 \alpha_2 \dots \alpha_{i-1} s_i$ . Thus,  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\text{Cyl}(\pi)) = 0$  if  $\pi$  is not a  $\mathfrak{S}$ -path. Given a measurable set  $\psi$  of maximal paths, then  $\text{Pr}_{\mathcal{M},s}^{\min}(\psi) = \inf_{\mathfrak{S}} \text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\psi)$  and  $\text{Pr}_{\mathcal{M},s}^{\max}(\psi) = \sup_{\mathfrak{S}} \text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\psi)$  where  $\mathfrak{S}$  ranges over all schedulers for  $\mathcal{M}$ . If  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\psi) > 0$  then the conditional probability measure  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\cdot | \psi)$  is given by  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\varphi | \psi) = \text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\varphi \cap \psi) / \text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\psi)$ . We write  $\text{Pr}_{\mathcal{M},s}^{\max}(\varphi | \psi)$  for the supremum of the values  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\varphi | \psi)$  where  $\mathfrak{S}$  ranges over all schedulers with  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\psi) > 0$ .

We often use LTL-like notations with the temporal modalities  $\bigcirc$  (next),  $\Diamond$  (eventually),  $\Box$  (always) and  $\text{U}$  (until) to specify measurable sets of maximal paths. For these it is well-known that optimal deterministic schedulers exists. If  $\psi$  is a reachability condition then even optimal deterministic memoryless schedulers exist. Let  $\emptyset \neq F \subseteq S$ . If  $\bowtie$  a comparison operator (e.g.  $=$  or  $\leq$ ) and  $r \in \mathbb{N}$  then  $\Diamond^{\bowtie r} F$  denotes the event “reaching  $F$  along some finite path  $\pi$  with  $\text{rew}(\pi) \bowtie r$ ”, while  $\bigcirc^{\bowtie n} F$  denotes “reaching  $F$  along some finite path  $\pi$  with  $|\pi| \bowtie n$ ”.

**(Conditional) expected rewards.** If  $F \subseteq S$  then  $\Diamond F$  denotes the random variable that assigns to each maximal path  $\varsigma$  in  $\mathcal{M}$  the reward  $\text{rew}(\pi)$  of the shortest prefix  $\pi$  of  $\varsigma$  where  $\text{last}(\pi) \in F$ . If  $\varsigma \not\models \Diamond F$  then  $(\Diamond F)(\varsigma) = \infty$ . If  $s \in S$  then  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F)$  denotes the expectation of  $\Diamond F$  in  $\mathcal{M}$  with starting state  $s$  under  $\mathfrak{S}$ , which is infinite if  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F) < 1$ .  $\mathbb{E}_{\mathcal{M},s}^{\max}(\Diamond F) \in \mathbb{R} \cup \{\pm\infty\}$  stands for  $\sup_{\mathfrak{S}} \mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F)$  where the supremum is taken over all schedulers  $\mathfrak{S}$  with  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F) = 1$  and  $\sup \emptyset = -\infty$ . If  $\psi$  is a measurable set of maximal paths and  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\psi) > 0$  then  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F | \psi)$  stands for the expectation of  $\Diamond F$  w.r.t. the conditional probability measure  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\cdot | \psi)$ .  $\mathbb{E}_{\mathcal{M},s}^{\max}(\Diamond F | \psi) \in \mathbb{R} \cup \{\pm\infty\}$  denotes the supremum of these conditional expectations when ranging over all schedulers  $\mathfrak{S}$  where  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\psi) > 0$  and  $\text{Pr}_{\mathcal{M},s}^{\mathfrak{S}}(\Diamond F | \psi) = 1$ .

**End components, MEC-quotient.** An *end component* of  $\mathcal{M}$  is a strongly connected sub-MDP. End components can be formalized as pairs  $\mathcal{E} = (E, \mathfrak{A})$

<sup>5</sup> Note that if  $\mathfrak{S}$  is reward-based then  $\mathfrak{S} \uparrow \pi = \mathfrak{S} \uparrow \pi'$  for all finite paths  $\pi, \pi'$  with  $(\text{last}(\pi), \text{rew}(\pi)) = (\text{last}(\pi'), \text{rew}(\pi'))$ .

where  $E$  is a nonempty subset of  $S$  and  $\mathfrak{A}$  a function that assigns to each state  $s \in E$  a nonempty subset of  $Act(s)$  such that the graph induced by  $\mathcal{E}$  is strongly connected.  $\mathcal{E}$  is called *maximal* if there is no end component  $\mathcal{E}' = (E', \mathfrak{A}')$  with  $\mathcal{E} \neq \mathcal{E}'$ ,  $E \subseteq E'$  and  $\mathfrak{A}(s) \subseteq \mathfrak{A}'(s)$  for all  $s \in E$ .  $\mathcal{E}$  is called *positive* if there exists a state-action pair  $(s, \alpha)$  with  $s \in E$ ,  $\alpha \in \mathfrak{A}(s)$  and  $rew(s, \alpha) > 0$ .

The *MEC-quotient* of an MDP  $\mathcal{M}$  is the MDP  $MEC(\mathcal{M})$  arising from  $\mathcal{M}$  by collapsing all states that belong to the same maximal end component [21]. Formally, we consider the equivalence relation  $\sim_{MEC}$  on the state space  $S$  of  $\mathcal{M}$  given by  $s \sim_{MEC} t$  iff  $s = t$  is not contained in some end component or  $s$  and  $t$  belong to the same maximal end component. Let  $[s]$  denote the equivalence class of state  $s$  with respect to  $\sim_{MEC}$ . The state space of  $MEC(\mathcal{M})$  is  $S / \sim_{MEC} = \{[s] : s \in S\}$ . The action set in  $MEC(\mathcal{M})$  is  $(S \times Act) \cup \{\tau\}$ . Action  $(s, \alpha)$  is enabled in state  $E \in S / \sim_{MEC}$  of  $MEC(\mathcal{M})$  iff  $s \in E$  and  $P(s, \alpha, t) > 0$  for at least one state  $t \in S \setminus E$ . In this case, the transition probabilities in  $MEC(\mathcal{M})$  are given by  $P_{MEC}(E, (s, \alpha), F) = P(s, \alpha, F)$ . If  $E$  is a bottom end component then  $E$  is trap state in  $MEC(\mathcal{M})$ . If  $G \subseteq S$  consists of trap states then the maximal probability to reach  $G$  in  $\mathcal{M}$  from  $s_{init}$  agrees with the maximal probability to reach  $G$  in  $MEC(\mathcal{M})$  (where we identify  $[s]$  and  $s$  if  $s$  is a trap). Reward functions can be lifted by  $rew_{MEC}(E, (s, \alpha)) = rew(s, \alpha)$ . However, for reasoning about reward-bounded properties the switch from  $\mathcal{M}$  to  $MEC(\mathcal{M})$  can cause problems. In general it is only justified if all state-action pairs that belong to some end component have reward 0.

## B Extremal conditional probabilities

We provide here a high-level overview of the reset-mechanism presented in [11] for the computation of maximal conditional probabilities for reachability objectives  $\Diamond F$  and conditions  $\Diamond G$  where  $F, G \subseteq S$ :

$$\Pr_{\mathcal{M}, s_{init}}^{\max}(\Diamond F | \Diamond G) = \sup_{\mathfrak{S}} \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond F | \Diamond G)$$

where  $\mathfrak{S}$  ranges over all schedulers with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond G) > 0$ . The approach of [11] uses a transformation of  $\mathcal{M}$  to a new MDP  $\mathcal{M}'$ . It relies on the observation that once  $G$  has been reached, the optimal behavior is to maximize reaching  $F$ . Similarly, once  $F$  has been reached, the optimal behavior is to maximize reaching  $G$ . This allows to capture the three relevant outcomes “goal” (both  $F$  and  $G$  have been seen), “stop” (“ $G$  but not  $F$  have been seen”) and “fail” (“ $G$  has not been seen”) by special trap states called *goal*, *stop* and *fail*. More precisely, transition to *fail* are inserted for all states  $s$  where  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond G) = 0$ . By adding a reset-transition with probability 1 from *fail* back to  $s_{init}$ , the probabilities for the paths that never visit  $G$  are “redistributed” to the paths that eventually enter  $G$ . This yields:

$$\Pr_{\mathcal{M}, s_{init}}^{\max}(\Diamond F | \Diamond G) = \Pr_{\mathcal{M}', s_{init}}^{\max}(\Diamond goal),$$

Thus, the computation of the maximal conditional probability for reachability objectives and conditions can be reduced to the computation of a maximal unconditional reachability probability, yielding a polynomial time bound.

The following example illustrates why the reset-mechanism presented in [11] is not adequate to compute maximal conditional expected accumulated rewards.

*Example B.1 (Reset-mechanism fails for conditional expectations).* Consider the Markov chain  $\mathcal{M}$  consisting of the initial state  $s = s_{init}$  and two trap states  $goal$  and  $fail$  with the transition probabilities  $P(s, goal) = P(s, fail) = \frac{1}{2}$ . Suppose the reward for state  $s$  is 1 and that  $F = G = \{goal\}$ .<sup>6</sup> Then, the expected accumulated reward for reaching  $goal$  under the condition to reach  $goal$  is simply the reward of the path  $\pi = s \ goal$ , which is 1. The reset-mechanism introduced for the computation of conditional reachability probabilities introduces a reset-transition from  $fail$  to  $s_{init}$ . For  $\psi = \Diamond goal$ , taking the reset-transition in the resulting Markov chain  $\mathcal{M}'$  corresponds to discarding all paths that eventually enter  $fail$  and redistributing their probabilities to the successful paths that eventually reach  $goal$ . Thus, the (only) successful  $\pi$  in  $\mathcal{M}$  is mimicked in  $\mathcal{M}'$  by the paths  $\pi_n = (s \ fail)^n s \ goal$ . The total probability of the (cylinder sets spanned by) paths  $\pi_n$  in  $\mathcal{M}'$  is 1, which agrees with the conditional probability of  $\pi$  in  $\mathcal{M}$ . However, the rewards of the paths  $\pi_n$  are different from  $rew(\pi)$ . Indeed we have:

$$E_{\mathcal{M}',s}(\Diamond goal) = \sum_{n=1}^{\infty} \left(\frac{1}{2}\right)^n \cdot n = 2 > 1 = E_{\mathcal{M},s}(\Diamond goal \mid \Diamond goal)$$

if we assign reward 0 to the reset-transition (resp. to state  $fail$ ). ■

## C Finiteness and upper bound

This section provides the proof for Theorem 1 and the details and soundness proofs for the methods to check finiteness of maximal conditional expectations and to compute an upper bound as outlined in Section 3. Throughout this section, we suppose that the given MDP  $\mathcal{M} = (S, Act, P, s_{init}, rew)$  has two distinguished sets  $F$  and  $G$  of states such that there is at least one scheduler  $\mathfrak{S}$  with  $\Pr_{\mathcal{M},s_{init}}^{\mathfrak{S}}(\Diamond G) > 0$  and  $\Pr_{\mathcal{M},s_{init}}^{\mathfrak{S}}(\Diamond F \mid \Diamond G) = 1$ . This condition can be checked in polynomial time using the reset-approach for conditional probabilities of [11] (see also Section B).

### C.1 Finiteness – preprocessing and normal form transformation

We now present the details of the preprocessing and normal form transformation. After some cleaning-up (step 1), we describe the transformations  $\mathcal{M} \rightsquigarrow \tilde{\mathcal{M}} \rightsquigarrow \tilde{\mathcal{M}}'$  (steps 2 and 3). Section C.2 will then transform  $\tilde{\mathcal{M}}'$  into the MDP  $\hat{\mathcal{M}}$  satisfying properties (1) and (2) presented in Section 3, which will then be subject for checking finiteness.

**Step 1: cleaning-up and assumptions.** Obviously, all states not reachable from  $s_{init}$  can be removed without affecting the conditional expected accumulated

<sup>6</sup> As  $\mathcal{M}$  is a Markov chain, action names for the transitions are irrelevant and can be omitted. Thus, the states of  $\mathcal{M}$  are decorated with reward values.

rewards from  $s_{init}$ . Thus, it is no restriction to suppose that all states  $s \in S$  are reachable from  $s_{init}$ . We can also safely assume that  $s_{init} \notin F \cup G$ . Note that  $s_{init} \in F$  would imply that the accumulated reward until  $F$  is 0 under each scheduler, while assumption  $s_{init} \in G$  would yield that  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond G) = 1$  for all schedulers  $\mathfrak{S}$ , in which case standard linear-programming techniques to compute (unconditional) maximal expected accumulated rewards can be applied.

**Step 2: normal form transformation.** We first show that there is a transformation  $\mathcal{M} \mapsto \tilde{\mathcal{M}}$  that permits to assume that  $F = G$ . Intuitively,  $\tilde{\mathcal{M}}$  operates in four modes: “normal mode”, “after  $G$ ”, “after  $F$ ” and “goal”.  $\tilde{\mathcal{M}}$  starts in normal mode where it behaves as  $\mathcal{M}$  as long as neither  $F$  nor  $G$  have been visited.

- If a  $G \setminus F$ -state  $u$  has been reached in normal mode then  $\tilde{\mathcal{M}}$  switches to the mode “after  $G$ ” where again it simulates  $\mathcal{M}$  and attempts to reach  $F$ .
- If an  $F \setminus G$ -state  $t$  has been reached in normal mode then  $\tilde{\mathcal{M}}$  switches to the mode “after  $F$ ” still simulating  $\mathcal{M}$  and expecting to reach a  $G$ -state.
- $\tilde{\mathcal{M}}$  enters the goal mode (consisting of a single trap state) as soon as a path fragment containing a state in  $F$  and a state in  $G$  has been generated, which is the case if  $\mathcal{M}$  visits an  $F$ -state in mode “after  $G$ ” or visits a  $G$ -state in mode “after  $F$ ”, or visits a state in  $F \cap G$  in the normal mode.

The rewards in the normal mode and in mode “after  $G$ ” are precisely as in  $\mathcal{M}$ , while the rewards are 0 in all other cases. The objective for  $\tilde{\mathcal{M}}$  is then to find a scheduler  $\mathfrak{T}$  such that  $\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\mathfrak{T}}(\Diamond goal)$  is positive and that optimizes the accumulated reward to reach the goal-state under the condition that the goal state will indeed be reached.

Formally, the state space of  $\tilde{\mathcal{M}}$  is  $\tilde{S} = S \cup S^G \cup S^F \cup \{goal\}$  where  $S^G$  and  $S^F$  consist of pairwise distinct copies of all states in  $\mathcal{M}$ . More generally, for  $U \subseteq S$ ,  $U^G = \{s^G : s \in U\}$  and  $U^F = \{s^F : s \in U\}$  with pairwise distinct, fresh states  $s^G$  and  $s^F$ .

The action set of  $\tilde{\mathcal{M}}$  is the action set  $Act$  of  $\mathcal{M}$  extended by a fresh action  $\tau$ . The transition probability function  $\tilde{P}$  and the reward function  $\tilde{rew} : \tilde{S} \times Act \rightarrow \mathbb{N}$  of  $\tilde{\mathcal{M}}$  are defined as follows. The new state  $goal$  is a trap. The transition probabilities for the normal mode are as follows (where  $v$  ranges over all states  $s \in S$ ):

- If  $s \in S \setminus (F \cup G)$  then  $Act_{\tilde{\mathcal{M}}}(s) = Act_{\mathcal{M}}(s)$ ,  $\tilde{P}(s, \alpha, v) = P(s, \alpha, v)$ ,  $\tilde{rew}(s, \alpha) = rew(s, \alpha)$ .
- If  $s \in F \cap G$  then  $Act_{\tilde{\mathcal{M}}}(s) = \{\tau\}$  and  $\tilde{P}(s, \tau, goal) = 1$ ,  $\tilde{rew}(s, \tau) = 0$ .
- The switches from normal mode to mode “after  $G$ ” resp. “after  $F$ ” are formalized as follows. If  $u \in G \setminus F$  and  $t \in F \setminus G$  and  $v \in S$  then:

$$\begin{aligned} \tilde{P}(u, \alpha, v^G) &= P(u, \alpha, v), & \tilde{rew}(u, \alpha) &= rew(u, \alpha), \\ \tilde{P}(t, \alpha, v^F) &= P(t, \alpha, v), & \tilde{rew}(t, \alpha) &= 0 \end{aligned}$$

and action  $\tau$  is not enabled in  $u$  or  $t$ , i.e.,  $Act_{\tilde{\mathcal{M}}}(s) = Act_{\mathcal{M}}(s)$  for all  $s \in (G \setminus F) \cup (F \setminus G)$ .

The transition probabilities and rewards in the modes “after  $G$ ” are defined as follows.

- If  $s \in S \setminus F$  then  $Act_{\tilde{\mathcal{M}}}(s^G) = Act_{\mathcal{M}}(s)$  and  $\tilde{P}(s^G, \alpha, v^G) = P(s, \alpha, v)$ ,  $\tilde{rew}(s^G, \alpha) = rew(s, \alpha)$ .
- If  $s \in F$  then  $Act_{\tilde{\mathcal{M}}}(s^G) = \{\tau\}$ ,  $\tilde{P}(s^G, \tau, goal) = 1$  and  $\tilde{rew}(s^G, \tau) = 0$ .

The transition probabilities and rewards in the mode “after  $F$ ” are defined analogously, except that the rewards for all state-action pairs in mode “after  $F$ ” are 0. That is:

- If  $s \in S \setminus G$  then  $Act_{\tilde{\mathcal{M}}}(s^F) = Act_{\mathcal{M}}(s)$ ,  $\tilde{P}(s^F, \alpha, v^F) = P(s, \alpha, v)$  and  $\tilde{rew}(s^F, \alpha) = 0$ .
- If  $s \in G$  then  $Act_{\tilde{\mathcal{M}}}(s^F) = \{\tau\}$ ,  $\tilde{P}(s^F, \tau, goal) = 1$  and  $\tilde{rew}(s^F, \tau) = 0$ .

Since the mode-switches in  $\mathcal{M}$  are deterministic, there is a one-to-one correspondence between the finite paths  $\pi$  in  $\mathcal{M}$  starting in  $s_{init}$  and visiting an  $F$ -state and a  $G$ -state (of minimal length with this property) and the finite paths in  $\tilde{\mathcal{M}}$  that lead from  $s_{init}$  to  $goal$ . This yields a transformation of a given scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  into a scheduler  $\tilde{\mathfrak{S}}$  for  $\tilde{\mathcal{M}}$  such that:

$$\mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond F | \Diamond G) = \mathbb{E}_{\tilde{\mathcal{M}}, s_{init}}^{\tilde{\mathfrak{S}}}(\Diamond goal | \Diamond goal)$$

Here, we suppose that  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond G) > 0$  and  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond F | \Diamond G) = 1$ . Vice versa, each scheduler  $\mathfrak{T}$  for  $\tilde{\mathcal{M}}$  with  $\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\mathfrak{T}}(\Diamond goal) > 0$  and either  $\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\mathfrak{T}}(\Diamond S^G) = 0$  or  $\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\mathfrak{T}}(\Diamond goal | \Diamond S^G) = 1$  induces a scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond G) > 0$  and  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond F | \Diamond G) = 1$  and such that  $\mathfrak{T} = \tilde{\mathfrak{S}}$ .

*Cleaning-up after Step 2.* The MDP  $\tilde{\mathcal{M}}$  can be further simplified using standard techniques without affecting the maximal condition expected accumulated reward. First, we simplify the sub-MDPs in the modes “after  $F$ ” and “after  $G$ ” where  $G$  resp.  $F$  will be reached almost surely under all schedulers as follows. We introduce a new action symbol  $\tau$ , discard the enabled actions in each state  $s^F$  where  $s \notin G$  and  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond G) = 1$  and add a new  $\tau$ -transition with reward 0 from  $s^F$  to  $goal$  with probability 1. Note that for such states  $s$  we have  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond G) = 1$  iff  $\Pr_{\mathcal{M}, s^F}^{\min}(\Diamond goal) = 1$ .

Likewise, for the states  $s^G$  with  $s \notin F$  and  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond F) = 1$  we can discard the enabled actions of  $s^G$ , while adding a  $\tau$ -transition from  $s^G$  to  $goal$  with probability 1. The reward of  $(s^G, \tau)$  is defined as the unconditional maximal expected accumulated reward to reach  $F$  from state  $s$  in  $\mathcal{M}$ ,

Finally, if  $U$  denotes the set of states  $u \in \tilde{S}$  with  $Act_{\tilde{\mathcal{M}}}(u) = \{\tau\}$ ,  $\tilde{P}(u, \tau, goal) = 1$  and  $\tilde{rew}(u, \tau) = 0$  then we can identify all states  $u \in U$  in  $\tilde{\mathcal{M}}$  with  $goal$ . Formally, the latter means that we replace  $\tilde{\mathcal{M}}$  with the MDP  $\hat{\mathcal{M}}$  arising from  $\tilde{\mathcal{M}}$  by removing all states  $u \in U$  and redefining the transition probability function by

$$\hat{P}(\tilde{s}, \alpha, goal) = \tilde{P}(\tilde{s}, \alpha, goal) + \sum_{u \in U} \tilde{P}(\tilde{s}, \alpha, u)$$



for all states  $\tilde{s} \in \tilde{S} \setminus U$  and all actions  $\alpha$ . The probabilities of all other transitions and the reward function remain unchanged. That is,  $\hat{P}(\tilde{s}, \alpha, \tilde{t}) = \tilde{P}(\tilde{s}, \alpha, \tilde{t})$  and  $\hat{r}ew(\tilde{s}, \alpha) = \tilde{r}ew(\tilde{s}, \alpha)$  for all states  $\tilde{s}, \tilde{t} \in \tilde{S} \setminus U$  and all actions  $\alpha$ .

**Step 3: auxiliary trap state.** We now perform a further transformation  $\tilde{\mathcal{M}} \rightsquigarrow \tilde{\mathcal{M}}'$  where  $\tilde{\mathcal{M}}'$  arises from the normal form MDP  $\tilde{\mathcal{M}}$  generated in Step 2. The new MDP  $\tilde{\mathcal{M}}'$  arises from  $\tilde{\mathcal{M}}$  by first removing all states  $\tilde{t}$  in the “after  $G$ ” mode with  $\Pr_{\tilde{\mathcal{M}}, \tilde{t}}^{\max}(\Diamond goal) < 1$ . Let  $V$  be the smallest set of states and state-action pairs that contains all states  $\tilde{t}$  in the “after  $G$ ” mode with  $\Pr_{\tilde{\mathcal{M}}, \tilde{t}}^{\max}(\Diamond goal) < 1$  and such that:

- for all states  $\tilde{v} \in V$ : if  $(\tilde{s}, \alpha)$  is a state-action pair in  $\tilde{\mathcal{M}}$  with  $\tilde{P}(\tilde{s}, \alpha, \tilde{v}) > 0$  then  $(\tilde{s}, \alpha) \in V$
- if  $(\tilde{s}, \beta) \in V$  for all  $\beta \in Act_{\tilde{\mathcal{M}}}(\tilde{s})$  and  $\tilde{s}$  is not a trap state then  $\tilde{s} \in V$ .

Let  $\tilde{\mathcal{M}}_0$  be the sub-MDP of  $\tilde{\mathcal{M}}$  consisting of the states  $\tilde{s} \in \tilde{S}$  with  $\tilde{s} \notin V$  and the action sets:

$$Act_{\tilde{\mathcal{M}}_0}(\tilde{s}) = Act_{\tilde{\mathcal{M}}}(\tilde{s}) \setminus \{\alpha : (\tilde{s}, \alpha) \in V\}$$

Then,  $\tilde{\mathcal{M}}'$  results from  $\tilde{\mathcal{M}}_0$  by:

- removing all states  $\tilde{t}$  where *goal* is not reachable from  $\tilde{t}$  in  $\tilde{\mathcal{M}}$  by redirecting all incoming transitions of  $\tilde{t}$  into *fail*,
- adding new transitions from the states  $\tilde{s}$  with  $\Pr_{\tilde{\mathcal{M}}, \tilde{s}}^{\min}(\Diamond goal) = 0$  to *fail*, provided  $\tilde{s}$  is not in the “after  $G$ ” mode.

We shall use the action label  $\iota$  for transitions to *fail*. More precisely, the state space of  $\tilde{\mathcal{M}}'$  is:  $\tilde{S}' = (\tilde{S}_0 \cup \{\text{fail}\}) \setminus T$  where  $\tilde{S}_0$  is the state space of  $\tilde{\mathcal{M}}_0$  and

$$T = \{ \tilde{t} \in \tilde{S}_0 : \Pr_{\tilde{\mathcal{M}}_0, \tilde{t}}^{\max}(\Diamond goal) = 0 \}$$

Note that  $s_{init} \notin T$  as we require  $\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\max}(\Diamond F | \Diamond G) = 1$ . The action set of  $\tilde{\mathcal{M}}'$  extends the action set  $Act_{\tilde{\mathcal{M}}}$  of  $\tilde{\mathcal{M}}$  by a fresh action  $\iota$ . Then, the transition probability function  $\tilde{P}'$  of  $\tilde{\mathcal{M}}'$  for the states  $\tilde{s}, \tilde{s}' \in \tilde{S}_0 \setminus T$  and actions  $\alpha \in Act_{\tilde{\mathcal{M}}}(\tilde{s})$  is given by:

$$\tilde{P}'(\tilde{s}, \alpha, \tilde{s}') = \tilde{P}(\tilde{s}, \alpha, \tilde{s}'), \quad \tilde{P}'(\tilde{s}, \alpha, \text{fail}) = \sum_{\tilde{t} \in T} \tilde{P}(\tilde{s}, \alpha, \tilde{t})$$

That is,  $\tilde{\mathcal{M}}'$  collapses all states in  $T$  into the single trap state *fail*. The reward of the state-action pairs  $(\tilde{s}, \alpha)$  with  $\tilde{s} \in \tilde{S}$  and  $\alpha \in Act_{\tilde{\mathcal{M}}}(\tilde{s})$  in  $\tilde{\mathcal{M}}'$  is the same as in  $\tilde{\mathcal{M}}$ . Finally, we add transitions from every state  $\tilde{s}$  in  $\tilde{\mathcal{M}}_0$  with  $\tilde{s} \notin S^G$  and  $\Pr_{\tilde{\mathcal{M}}, \tilde{s}}^{\min}(\Diamond goal) = 0$  to *fail* with action label  $\iota$ . That is, in all those states  $\tilde{s}$ ,  $\iota$  is an additional enabled action with the transition probability  $\tilde{P}'(\tilde{s}, \iota, \text{fail}) = 1$  and reward 0 for the state-action pair  $(\tilde{s}, \iota)$ .

**Lemma C.1 (Soundness of the transformation).** *Let  $\mathcal{M}$  denote the original MDP and  $\tilde{\mathcal{M}}'$  the MDP resulting from the transformations in steps 1,2 and 3. Then:*

$$\mathbb{E}_{\mathcal{M}, s_{init}}^{\max} (\Diamond F \mid \Diamond G) = \sup_{\mathfrak{T}'} \mathbb{E}_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond goal \mid \Diamond goal)$$

where the supremum on the right ranges over all schedulers  $\mathfrak{T}'$  for  $\tilde{\mathcal{M}}'$  such that  $\Pr_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond goal) > 0$  and  $\Pr_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond(goal \vee fail)) = 1$ .

Note that this does not imply the finiteness of  $\mathbb{E}_{\mathcal{M}, s_{init}}^{\max} (\Diamond F \mid \Diamond G)$ , which will be checked with the methods of the next section (Section C.2).

*Proof.* Recall that  $\mathbb{E}_{\mathcal{M}, s_{init}}^{\max} (\Diamond F \mid \Diamond G)$  has been defined as the supremum of the conditional expectations  $\mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond F \mid \Diamond G)$  where  $\mathfrak{S}$  ranges over all schedulers for  $\mathcal{M}$  such that  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond G) > 0$  and  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond F \mid \Diamond G) = 1$ . Let  $Sched(\mathcal{M}, F, G)$  denote this class of schedulers.

Let  $Sched(\tilde{\mathcal{M}}', goal)$  denote the class of schedulers  $\mathfrak{T}'$  for  $\tilde{\mathcal{M}}'$  such that  $\Pr_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond goal) > 0$  and  $\Pr_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond(goal \vee fail)) = 1$ .

Each scheduler  $\mathfrak{S}$  in  $Sched(\mathcal{M}, F, G)$  naturally induces a scheduler  $\mathfrak{T}'$  in  $Sched(\tilde{\mathcal{M}}', goal)$  such that  $\mathfrak{T}'$  mimics the  $\mathfrak{S}$ -paths satisfying  $\Diamond F \wedge \Diamond G$  by  $\mathfrak{T}'$ -paths to  $goal$  and such that:

$$\mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond F \mid \Diamond G) = \mathbb{E}_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond goal \mid \Diamond goal)$$

To see this, let  $\mathfrak{T}$  denote the lifting of  $\mathfrak{S}$  to a scheduler for  $\tilde{\mathcal{M}}$ . Then:

$$\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\mathfrak{T}} (\Diamond goal) = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond F \wedge \Diamond G) > 0$$

Scheduler  $\mathfrak{T}'$  for  $\tilde{\mathcal{M}}'$  behaves as  $\mathfrak{T}$  for all finite paths  $\tilde{\pi}$  where  $\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\mathfrak{T} \uparrow \tilde{\pi}} (\Diamond goal) > 0$ . As soon as  $\mathfrak{T}'$  has generated a path  $\tilde{\pi}$  with  $\Pr_{\tilde{\mathcal{M}}, s_{init}}^{\mathfrak{T} \uparrow \tilde{\pi}} (\Diamond goal) = 0$ , then  $\mathfrak{T}'$  schedules action  $\iota$ . Then,  $\mathfrak{T}$  and  $\mathfrak{T}'$  have the same paths from  $s_{init}$  to  $goal$ , and these correspond to the  $\mathfrak{S}$ -paths satisfying  $\Diamond F \wedge \Diamond G$ . The probabilities and accumulated rewards of these paths in  $\tilde{\mathcal{M}}'$ ,  $\tilde{\mathcal{M}}$  and  $\mathcal{M}$  are the same. This yields:

$$\mathbb{E}_{\mathcal{M}, s_{init}}^{\max} (\Diamond F \mid \Diamond G) \leq \sup_{\mathfrak{T}'} \mathbb{E}_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond goal \mid \Diamond goal)$$

Vice versa, we show that for each  $\mathfrak{T}'$  in  $Sched(\tilde{\mathcal{M}}', goal)$  there exists a scheduler  $\mathfrak{S}$  in  $Sched(\mathcal{M}, F, G)$  such that:

$$\mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}} (\Diamond F \mid \Diamond G) \geq \mathbb{E}_{\tilde{\mathcal{M}}', s_{init}}^{\mathfrak{T}'} (\Diamond goal \mid \Diamond goal)$$

Let us first suppose that  $\tilde{\mathcal{M}}$  and  $\tilde{\mathcal{M}}'$  have a positive end component  $\tilde{\mathcal{E}}$  in the “after  $G$ ” mode such that  $goal$  is reachable from  $\tilde{\mathcal{E}}$  and  $\tilde{\mathcal{E}}$  is reachable from  $s_{init}$ . We show that in this case the maximal conditional expectation of  $\mathcal{M}$  is infinite. Let  $\mathfrak{U}$  be a scheduler that “realizes” this end component. That is, if  $\psi$  denotes

the event “take any state-action pair of  $\mathcal{E}$  infinitely often” then  $\Pr_{\mathcal{M}, \tilde{t}}^{\mathfrak{U}}(\psi) = 1$  for each state  $\tilde{t}$  of  $\tilde{\mathcal{E}}$ . Obviously,  $\tilde{\mathcal{E}}$  corresponds to some positive end component of  $\mathcal{M}$  and  $\mathfrak{U}$  can also be viewed as a scheduler for  $\mathcal{M}$  that “realizes”  $\mathcal{E}$ . Furthermore, we pick some memoryless scheduler  $\mathfrak{V}$  for  $\mathcal{M}$  and a shortest path  $\varrho$  in  $\mathcal{M}$  from  $s_{init}$  to  $\mathcal{E}$  as well as a shortest path  $\pi$  from  $\mathcal{E}$  to some state in  $F^G$ . (Such a path exists as *goal* is reachable from  $\mathcal{E}$  in  $\tilde{\mathcal{M}}$ .) Let now  $R \in \mathbb{N}$  and let  $\mathfrak{S}_R$  be the following scheduler. In its first mode,  $\mathfrak{S}_R$  attempts to generate the path  $\varrho$ . If it fails then it switches mode and behaves as  $\mathfrak{V}$  from then on. As soon as  $\varrho$  has been generated then  $\mathfrak{S}_R$  behaves as  $\mathfrak{U}$  as long as the accumulated rewards is smaller than  $R$ . As soon as the accumulated reward exceeds  $R$  then  $\mathfrak{S}_R$  switches mode again, leaves  $\mathcal{E}$  and attempts to reach  $F^G$  along  $\pi$ . If it fails then it behaves as  $\mathfrak{V}$ . For the conditional expectation of  $\mathfrak{S}_R$  we have:

$$\mathbb{CE}^{\mathfrak{S}_R}(\Diamond F | \Diamond G) \geq \frac{\rho + p \cdot R}{x + p}$$

where  $\rho, x \geq 0$  and  $p = \text{prob}(\varrho) \cdot \text{prob}(\pi)$ . The value  $x$  stands for the probability under  $\mathfrak{S}_R$  to reach *goal* via paths that do not have the form  $\varrho; \xi; \pi$  where  $\xi$  is a  $\mathfrak{U}$ -path, and  $\rho$  stands for the corresponding partial expectation. Thus, there exists some  $R \in \mathbb{N}$  with

$$\mathbb{CE}^{\mathfrak{S}_R}(\Diamond F | \Diamond G) \geq \mathbb{CE}^{\mathfrak{T}'}(\Diamond \text{goal} | \Diamond \text{goal})$$

Let us now suppose that there is no positive end component in the “after  $G$ ” mode of  $\tilde{\mathcal{M}}'$  that is reachable from  $s_{init}$  and from which *goal* is reachable. Let  $\mathfrak{T}' \in \text{Sched}(\tilde{\mathcal{M}}', \text{goal})$ . Let  $\tilde{U}$  denote the set of states  $\tilde{u}$  in  $\tilde{\mathcal{M}}$  with  $\Pr_{\tilde{\mathcal{M}}, \tilde{u}}^{\min}(\Diamond \text{goal}) = 0$ . Thus,  $\tilde{u} \in \tilde{U}$  iff  $\iota \in \text{Act}_{\tilde{\mathcal{M}}'}(\tilde{u})$ . Then,  $\tilde{U} \cap S^G = \emptyset$  by definition of  $\tilde{\mathcal{M}}'$ . We pick a memoryless scheduler  $\mathfrak{U}$  for  $\tilde{\mathcal{M}}$  such that  $\Pr_{\tilde{\mathcal{M}}, \tilde{u}}^{\mathfrak{U}}(\Diamond \text{goal}) = 0$  for all states  $\tilde{u} \in \tilde{U}$ . Then,  $\mathfrak{U}$  can also be viewed as a scheduler for  $\mathcal{M}$ . Scheduler  $\mathfrak{S}$  for an input path  $\pi$  in  $\tilde{\mathcal{M}}$  behaves as  $\mathfrak{T}'$  for the corresponding path  $\tilde{\pi}$  in  $\tilde{\mathcal{M}}'$  provided that  $\mathfrak{T}'(\tilde{\pi}) \neq \iota$ . As soon as  $\mathfrak{T}'$  schedules  $\iota$  then  $\tilde{u} \stackrel{\text{def}}{=} \text{last}(\tilde{\pi}) \in \tilde{U}$ . Then,  $\mathfrak{S}$  behaves as  $\mathfrak{U}$  from then on. Up to the mode-annotations of the states in  $\tilde{\mathcal{M}}'$ ,  $\mathfrak{S}$  and  $\mathfrak{T}'$  have the same “successful” paths (i.e., satisfying  $\Diamond F \wedge \Diamond G$  resp.  $\Diamond \text{goal}$ ) with the same probabilities and rewards. In particular, this yields  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond G) > 0$ . Furthermore, we have  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond F | \Diamond G) = 1$ . The latter is a consequence of the fact that  $\tilde{U} \cap S^G = \emptyset$ . This yields  $\mathfrak{S} \in \text{Sched}(\mathcal{M}, F, G)$  and that  $\mathfrak{S}$  and  $\mathfrak{T}'$  have the same maximal conditional expectations. ■

In summary, with the three preprocessing steps, we can transform  $\mathcal{M}$  into an MDP  $\tilde{\mathcal{M}}'$  with two trap states *goal* and *fail* such that  $s_{init} \notin \{\text{goal}, \text{fail}\}$ . In the next section, we will describe a further transformation MDP  $\tilde{\mathcal{M}}' \rightsquigarrow \hat{\mathcal{M}}$  such that  $\hat{\mathcal{M}}$  satisfies conditions (1) and (2) of Section 3.  $\hat{\mathcal{M}}$  will then be used for deciding finiteness (Section C.3), computing an upper bound  $\mathbb{CE}^{\text{ub}}$  for the  $\mathbb{CE}^{\text{max}}$  (Section C.4) and the subsequent threshold algorithm and the computation of an optimal scheduler as outlined in Section 4.

## C.2 Finiteness – critical schedulers

In the sequel, we suppose that  $\mathcal{M}$  is the result of the preprocessing and normal form transformation presented in the previous subsection. Thus, the task is to compute the maximal expected accumulated reward until reaching the trap state *goal* under the condition that *goal* will be reached where the supremum is taken over all schedulers  $\mathfrak{S}$  satisfying the following requirement (SR) (see Lemma C.1):

$$\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0 \quad \text{and} \quad \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond(goal \vee fail)) = 1 \quad (\text{SR})$$

Furthermore, for all states  $s$  in  $\mathcal{M}$ :

$$s \not\models \exists \Diamond goal \quad \text{iff} \quad s = fail \quad (\text{A1})$$

Before we start into this section, we will briefly repeat and summarize notations relevant for this section.

**Definition C.2 (Shortform notations for (conditional) expectations).** As before, we often write  $\Pr_s^{\mathfrak{S}}$  for  $\Pr_{\mathcal{M}, s}^{\mathfrak{S}}$ . If  $\mathfrak{S}$  is a scheduler for  $\mathcal{M}$  with  $\Pr_{s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$  then we shortly write  $\mathbb{CE}^{\mathfrak{S}}$  for the conditional expected accumulated reward until reaching the goal state under scheduler  $\mathfrak{S}$  under the condition that the goal state will indeed be reached. That is:

$$\mathbb{CE}^{\mathfrak{S}} = \mathbb{E}_{s_{init}}^{\mathfrak{S}}(\Diamond goal \mid \Diamond goal)$$

We often refer to  $\mathbb{CE}^{\mathfrak{S}}$  as the conditional expectation under  $\mathfrak{S}$ . Furthermore, let

$$\mathbb{CE}^{\max} = \sup_{\mathfrak{S}} \mathbb{CE}^{\mathfrak{S}} \quad \text{and} \quad \mathbb{CE}^{\min} = \inf_{\mathfrak{S}} \mathbb{CE}^{\mathfrak{S}}$$

where  $\mathfrak{S}$  ranges over all schedulers for  $\mathcal{M}$  satisfying the scheduler requirement (SR). We also often use the notation  $\mathbb{E}_s^{\mathfrak{S}}$  as a shortform for

$$\mathbb{E}_s^{\mathfrak{S}} = \mathbb{E}_{\mathcal{M}, s}^{\mathfrak{S}} = \sum_{r=0}^{\infty} r \cdot \Pr_{\mathcal{M}, s}^{\mathfrak{S}}(\Diamond^{\neg r} goal)$$

Here,  $\mathfrak{S}$  is an arbitrary scheduler in  $\mathcal{M}$  and  $s$  a state of  $\mathcal{M}$ . Clearly, we then have

$$\mathbb{CE}^{\mathfrak{S}} = \frac{\mathbb{E}_{s_{init}}^{\mathfrak{S}}}{\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal)}$$

for each scheduler  $\mathfrak{S}$  satisfying (SR). If  $s \in S \setminus \{goal, fail\}$  and  $\mathfrak{S}$  satisfies (SR) then:

$$\mathbb{CE}_s^{\mathfrak{S}} = \mathbb{E}_s^{\mathfrak{S}}(\Diamond goal \mid \Diamond goal) = \frac{\mathbb{E}_s^{\mathfrak{S}}}{\Pr_s^{\mathfrak{S}}(\Diamond goal)}$$

Hence,  $\mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}_{s_{init}}^{\mathfrak{S}}$ . ■

We will now present a criterion to decide whether  $\mathbb{CE}^{\max} < \infty$  and consider the unconditional case first.

*Remark C.3 (Unconditional maximal expected accumulated reward).* It is well-known that if  $\Pr_{\mathcal{M}, s_{init}}^{\min}(\Diamond goal) = 1$  then the unconditional expected accumulated reward  $E_{s_{init}}^{\mathfrak{S}} = \mathbb{E}_{s_{init}}^{\mathfrak{S}}(\Diamond goal)$  is finite for all schedulers  $\mathfrak{S}$ . Furthermore, there exists a memoryless deterministic scheduler maximizing the unconditional expected accumulated reward. In particular, the supremum of the unconditional expected accumulated rewards under all schedulers is finite.

However, if  $\Pr_{\mathcal{M}, s_{init}}^{\min}(\Diamond goal) < 1$  then the expected accumulated reward to reach *goal* can be infinite for (infinite-memory) schedulers  $\mathfrak{S}$  with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal) = 1$ . To illustrate this phenomenon, consider an MDP  $\mathcal{M}$  with two states  $s = s_{init}$  and *goal*, the transition probabilities  $P(s, \alpha, goal) = P(s, \beta, s) = 1$  and the reward  $rew(s, \alpha) = rew(s, \beta) = 1$ . We pick an infinite sequence  $(q_n)_{n \geq 1}$  of rational numbers in  $]0, 1[$  such that  $\sum_n q_n = 1$  and  $\sum_n n \cdot q_n$  diverges (e.g.,  $q_n = x/n^2$  where  $1/x$  is the value of the series  $\sum_n 1/n^2$ ). Furthermore, we put  $x_1 = 1$  and  $x_n = p_1 \cdot \dots \cdot p_{n-1}$  for  $n > 1$ . Let  $\mathfrak{S}$  be the randomized scheduler for  $\mathcal{M}$  that schedules action  $\beta$  with probability  $p_n = 1 - q_n/x_n$  and action  $\alpha$  with probability  $q_n/x_n$  for the  $n$ -th visit of state  $s$ . Let  $\pi_n$  be the path  $(s\beta)^{n-1}s\alpha goal$ . The probability of  $\pi_n$  under  $\mathfrak{S}$  is  $p_1 \cdot p_2 \cdot \dots \cdot p_{n-1} \cdot (1 - p_n) = x_n(1 - p_n) = q_n$ . Hence:

$$\Pr_{\mathcal{M}, s}^{\mathfrak{S}}(\Diamond goal) = \sum_{n=1}^{\infty} q_n = 1$$

Since  $rew(\pi_n) = n$ , the expected accumulated reward under  $\mathfrak{S}$  for reaching *goal* is  $\sum_n n \cdot x_n \cdot (1 - p_n) = \sum_n n \cdot q_n = \infty$ . ■

Let *PosEC* be the set of all states  $s$  in  $\mathcal{M}$  that belong to some end component  $\mathcal{E} = (E, \mathfrak{A})$  with  $\{goal, fail\} \cap E = \emptyset$  and  $rew(t, \alpha) \geq 1$  for some state-action pair  $(t, \alpha)$  with  $t \in E$  and  $\alpha \in \mathfrak{A}(t)$ . Such end components are said to be *positive*.

**Lemma C.4.** *If  $PosEC \neq \emptyset$  then  $\mathbb{CE}^{\max} = \infty$ .*

*Proof.* By assumption (A1) we have  $t \models \exists \Diamond goal$  for all  $t \in PosEC$ . Furthermore, for each infinite path  $\varsigma$ :  $\varsigma \models \Diamond PosEC$  iff  $\varsigma \models (\neg goal) \cup PosEC$ .

Let  $R \in \mathbb{N}$ . We pick some state  $s \in PosEC$  and schedulers  $\mathfrak{S}_s$  and  $\mathfrak{S}_g$  with

$$p_s = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}_s}(\Diamond s) > 0 \quad \text{and} \quad p_g = \Pr_s^{\mathfrak{S}_g}(\Diamond goal) > 0.$$

Furthermore, let  $\mathfrak{S}_{EC}$  be a scheduler such that the limit of almost all  $\mathfrak{S}_{EC}$ -paths starting in  $s$  is a positive end component containing  $s$ . We construct a scheduler  $\mathfrak{T}_R$  as follows. For input paths starting in  $s_{init}$ , scheduler  $\mathfrak{T}_R$  first behaves as  $\mathfrak{S}_s$  until state  $s$  has been reached (this happens with probability  $p_s$ ). It then switches its mode and behaves as  $\mathfrak{S}_{EC}$  until a path fragment  $\pi$  has been generated such that  $last(\pi) = s$  and  $rew(\pi) > R$  (this happens with probability 1). Having generated such a path fragment  $\pi$ ,  $\mathfrak{T}_R$  switches its mode again and behaves as  $\mathfrak{S}_g$  from then on. Then, the expected accumulated reward to reach *goal* under  $\mathfrak{T}_R$  has the form:

$$\mathbb{CE}^{\mathfrak{T}_R} \geq \frac{\rho + p_s \cdot p_g \cdot R}{x + p_s \cdot p_g}$$

where

$$\rho = \sum_{r=0}^{\infty} r \cdot \Pr_{s_{init}}^{\mathfrak{S}_s}(\neg s \text{ U}^{\neg r} \text{ goal}) \quad \text{and} \quad x = \Pr_{s_{init}}^{\mathfrak{S}_s}(\neg s \text{ U } \text{ goal})$$

Since  $\rho$  and  $x$  do not depend on  $R$ , we have:

$$\lim_{R \rightarrow \infty} \mathbb{CE}^{\mathfrak{T}_R} = \infty$$

Thus,  $\mathbb{CE}^{\max} = \infty$  if  $PosEC$  is nonempty.  $\blacksquare$

Obviously,  $PosEC = \emptyset$  if there is no positive maximal end component. Hence, the criterion of Lemma C.4 can be checked in time polynomial in the size of the MDP  $\mathcal{M}$  using the same techniques as proposed by de Alfaro [23] for computing maximal unconditional expectations.<sup>7</sup> In what follows, we suppose that  $\mathcal{M}$  has no positive end component. That is,  $rew(s, \alpha) = 0$  for all state-action pairs  $(s, \alpha)$  that belong to some (maximal) end component. But then we can collapse all maximal end components into a single state and discard all actions  $\alpha \in Act(s)$  where  $(s, \alpha)$  belongs to an end component (i.e., we identify  $\mathcal{M}$  with its MEC-quotient, see Appendix A), without affecting the maximal expected accumulated reward.

**Lemma C.5 (Maximal unconditional expectations (see [23])).** *Suppose  $PosEC = \emptyset$ . Then,  $\mathbb{E}_{\mathcal{M}, s_{init}}^{\max}(\Diamond \text{ goal}) < \infty$  and there exists a memoryless deterministic scheduler  $\mathfrak{S}$  with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond \text{ goal}) = 1$  and  $\mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond \text{ goal}) = \mathbb{E}_{\mathcal{M}, s_{init}}^{\max}(\Diamond \text{ goal})$ .*

We now return to the case of conditional expected accumulated rewards. Assuming  $PosEC = \emptyset$ , the transformation that collapses all maximal end components into a single state (see above) permits the following additional assumption:

$$\mathcal{M} \text{ has no end component} \tag{A2}$$

Under assumption (A2) we have  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond(\text{goal} \vee \text{fail})) = 1$  for all states  $s$ . Hence, the scheduler requirement (SR) reduces to  $\Pr_{\mathcal{M}, s}^{\mathfrak{S}}(\Diamond \text{ goal}) > 0$ . Note that after this transformation the MDP  $\mathcal{M} = \hat{\mathcal{M}}$  satisfies conditions (1) and (2) presented in Section 3.

The nonemptiness of  $PosEC$  is, however, not sufficient to cover all cases where the conditional expected accumulated reward is infinite. This is illustrated in the following example.

*Example C.6.* Let  $\mathcal{M}$  be the MDP  $\mathcal{M}[\mathfrak{r}]$  shown in Figure 1, but with initial state  $s_{init} = s_2$ . The parameter  $\mathfrak{r}$  is of no concern for this example. For the scheduler  $\mathfrak{S}_n$  that chooses action  $\beta$  for the first  $n$  visits of  $s$  and then  $\alpha$ , there is a single  $\mathfrak{S}_n$ -path reaching *goal*, namely  $\pi_n = (s \beta)^n s \alpha \text{ goal}$ . The accumulated reward of  $\pi_n$  is  $n$ , and so is the conditional expectation of  $\mathcal{M}$  under  $\mathfrak{S}_n$ . Thus,  $\mathbb{CE}^{\max}$  is infinite, although  $PosEC$  is empty.  $\blacksquare$

<sup>7</sup> More precisely, Section 4 of [23] addresses the computation of minimal expected accumulated reward in non-positive MDPs where the minimum is taken over all schedulers that reach the goal state almost surely.

**Definition C.7 (Positive, zero-reward, simple cycle; critical scheduler).**

Let  $\mathcal{M}$  be an MDP as before satisfying (A1) and (A2). A cyclic path  $\xi = t_0 \beta_0 t_1 \beta_1 \dots \beta_{n-1} t_n$  in  $\mathcal{M}$  is said to be *positive* if  $\text{rew}(t_i, \beta_i) > 0$  for at least one index  $i \in \{0, 1, \dots, n-1\}$ . Otherwise  $\xi$  is called a *zero-reward* cycle.  $\xi$  is said to be *simple* if  $t_i \neq t_j$  for  $0 \leq i < j < n$ .

Scheduler  $\mathfrak{U}$  is called *critical* if  $\Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{U}}(\Box \neg \text{goal}) = 1$  and there is a reachable positive  $\mathfrak{U}$ -cycle, i.e., there is a finite  $\mathfrak{U}$ -path  $s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{m-1} s_m$  starting in  $s_0 = s_{\text{init}}$  such that for some  $k < m$  the suffix  $s_k \alpha_k s_{k+1} \alpha_{k+1} \dots \alpha_{m-1} s_m$  is a positive cycle. ■

Note that  $\Pr_{s_{\text{init}}}^{\mathfrak{U}}(\Diamond \text{fail}) = 1$  for each critical scheduler by assumption (A2). Thus, if  $\Pr_s^{\min}(\Diamond \text{goal}) > 0$  for all states  $s \in S \setminus \{\text{fail}\}$  then  $\mathcal{M}$  has no critical scheduler.

**Proposition C.8 (Infinite maximal conditional expected rewards).**

With the notations and assumptions (A1) and (A2) as above, the following three statements are equivalent:

- (i)  $\mathbb{CE}^{\max} = \infty$
- (ii)  $\mathcal{M}$  has a memoryless deterministic critical scheduler
- (iii)  $\mathcal{M}$  has a critical scheduler

*Proof.* We first show the equivalence of statements (ii) and (iii). The implication (ii)  $\implies$  (iii) is trivial. For the implication (iii)  $\implies$  (ii), we suppose that we are given a (possibly randomized history-dependent) critical scheduler  $\mathfrak{U}$ . A memoryless deterministic critical scheduler  $\mathfrak{U}'$  is obtained by picking a simple positive  $\mathfrak{U}$ -cycle  $\xi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{n-1} s_n$ . We then put  $\mathfrak{U}'(s_i) = \alpha_i$  for  $0 \leq i < n$ . For each state  $s \in S \setminus \{s_0, \dots, s_{n-1}\}$  we pick an arbitrary action  $\alpha$  with  $\mathfrak{U}(s)(\alpha) > 0$  and define  $\mathfrak{U}'(s) = \alpha$ .

(ii)  $\implies$  (i): Suppose  $\mathcal{M}$  has a memoryless deterministic critical scheduler  $\mathfrak{U}$ . The argument is similar to the proof of Lemma C.4. Let  $\xi = t_0 \beta_0 t_1 \beta_1 \dots \beta_{n-1} t_n$  be positive  $\mathfrak{U}$ -cycle and  $s = t_0 = t_n$ . Furthermore, let  $\mathfrak{S}$  be a memoryless deterministic scheduler under which  $s$  reaches the goal-state with positive probability. Then,  $p_g = \Pr_s^{\mathfrak{S}}(\Diamond \text{goal}) > 0$ . Let  $p_s = \Pr_{s_{\text{init}}}^{\mathfrak{U}}(\Diamond s)$  and  $p_c = \Pr_s^{\mathfrak{U}}(\text{Cyl}(\xi))$ . Then,  $p_s > 0$  and  $p_c = \text{prob}(\xi) > 0$ . For  $R \in \mathbb{N}$ , we construct a scheduler  $\mathfrak{T}_R$  with two modes as follows.

- In its first mode,  $\mathfrak{T}_R$  behaves as the critical scheduler  $\mathfrak{U}$  for all input paths  $\pi$  where  $\xi^R$  is not a fragment of  $\pi$ .
- If the input path  $\pi$  has a suffix that runs  $R$ -times through the cycle  $\xi$  and has not visited state  $s$  before entering  $\xi$  (in which case  $\text{last}(\pi) = s$  and  $\text{rew}(\pi) \geq R$ ), then  $\mathfrak{T}_R$  switches mode and behaves as  $\mathfrak{S}$  from now on.

Note that the switch from the first mode (simulation of  $\mathfrak{U}$ ) to the second mode (simulation of  $\mathfrak{S}$ ) appears with probability  $p_s \cdot p_c^R$ . As  $\Pr_{s_{\text{init}}}^{\mathfrak{U}}(\Box \neg \text{goal}) = 1$ , we

have  $\Pr_{s_{init}}^{\mathfrak{T}_R}(\neg s \text{ U } goal) = 0$ . For the conditional expected accumulated reward to reach *goal* under  $\mathfrak{T}_R$ , we get:

$$\mathbb{CE}^{\mathfrak{T}_R} \geq \frac{p_s \cdot p_c^R \cdot p_g \cdot R}{p_s \cdot p_c^R \cdot p_g} = R$$

Hence, the limit of  $\mathbb{CE}^{\mathfrak{T}_R}$  is infinite if  $R$  tends to infinity.

(i)  $\implies$  (iii): We now suppose that  $\mathcal{M}$  has no critical scheduler and show that the conditional expected accumulated reward is finite. For this, we adapt the reset-mechanism that has been introduced for computing conditional reachability probabilities in [11] (see Appendix B) and rely on Lemma C.5.

The reset-mechanism is, however, not directly applicable since the resulting MDP  $\mathcal{N}$  with reset-transitions from the fail-state to the initial state might have positive end components, in which case the maximal unconditional expected accumulated reward can be infinite. For this reason, we first transform  $\mathcal{M}$  into a new MDP  $\mathcal{M}'$  that has the same probabilistic structure and the same schedulers, but uses a different reward structure. The maximal conditional expected accumulated rewards to reach *goal* are the same in  $\mathcal{M}$  and  $\mathcal{M}'$ . Having constructed  $\mathcal{M}'$  we then can implement the reset-mechanism and switch from  $\mathcal{M}'$  to a new MDP  $\mathcal{N}$  that has no positive end component. The maximal unconditional expected accumulated reward to reach *goal* in  $\mathcal{N}$  is finite (by Lemma C.5) and yields an upper bound  $\mathbb{CE}^{\text{ub}}$  for the maximal conditional expected accumulated reward to reach *goal* in  $\mathcal{M}$  (or  $\mathcal{M}'$ ).

*The new MDP  $\mathcal{M}'$ .* We first provide an informal explanation of the behavior of the MDP  $\mathcal{M}'$ . The idea is that  $\mathcal{M}'$  behaves as  $\mathcal{M}$ , but postpones the assignment of rewards until it is clear that *goal* will be reached with positive probability. Note that we can deal with  $\mathcal{M}' = \mathcal{M}$  if  $\Pr_s^{\min}(\Diamond goal) > 0$  for all states  $s \neq fail$ . The following construction only serves to deal with the case where  $\mathcal{M}$  contains some non-trap states  $s$  with  $\Pr_s^{\min}(\Diamond goal) = 0$ .

The definition of  $\mathcal{M}'$  relies on the following observation. Let

$$R = \sum_{s \in S} rew^{\max}(s)$$

where  $rew^{\max}(s) = 0$  if  $s = goal$  or  $s = fail$  and  $rew^{\max}(s) = \max\{rew(s, \alpha) : \alpha \in Act(s)\}$  for  $s \in S \setminus \{goal, fail\}$ . Clearly, if  $\pi$  is a finite path in  $\mathcal{M}$  with  $rew(\pi) > R$  then  $\pi$  contains a positive cycle. As  $\mathcal{M}$  has no critical schedulers, we have  $\Pr_{s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$  for each scheduler  $\mathfrak{S}$  where  $rew(\pi) > R$  for some  $\mathfrak{S}$ -path  $\pi$  starting in  $s_{init}$ .

The new MDP  $\mathcal{M}'$  simulates  $\mathcal{M}$  while operating in two modes. In its first mode,  $\mathcal{M}'$  augments the states with the information on the reward that has been accumulated in the past. That is, the states of  $\mathcal{M}'$  in its first mode are pairs  $\langle s, r \rangle \in S \times \mathbb{N}$  where  $r = rew(\pi) \leq R$  for some finite path  $\pi$  in  $\mathcal{M}$  from  $s_{init}$  to  $s$ . Thus, as soon as  $\mathcal{M}'$  takes an action  $\alpha$  in state  $\langle s, r \rangle$  where  $r + rew(s, \alpha)$  exceeds  $R$  then  $\mathcal{M}'$  switches to the second mode where it behaves exactly as  $\mathcal{M}$  without any reward annotations of the states. We assign reward 0 to all state-action pairs



in the first mode. The reward of the switches from the first to the second mode, say from state  $\langle s, r \rangle$  via firing action  $\alpha$ , is the total reward accumulated so far (value  $r$ ) plus the reward of the taken action (the value  $rew(s, \alpha)$ ). The rewards for the state-action pairs in the second mode are the same as in  $\mathcal{M}$ .

The definition of the new MDP  $\mathcal{M}'$  is as follows. The state space of  $\mathcal{M}'$  is:

$$S' = S \times \{0, 1, \dots, R\} \cup S$$

The action set of  $\mathcal{M}'$  is  $Act' = Act$ . The computation of  $\mathcal{M}'$  starts in state  $s'_{init} = \langle s_{init}, 0 \rangle$ . The transition probability function  $P'$  and reward function  $rew'$  are defined as follows.

- Let  $s \in S \setminus \{goal\}$ ,  $r \in \{0, 1, \dots, R\}$ ,  $\alpha \in Act(s)$  and  $r' = r + rew(s, \alpha)$ .
  - If  $r' \leq R$  then  $P'(\langle s, r \rangle, \alpha, \langle t, r' \rangle) = P(s, \alpha, t)$  for all states  $t \in S$  and  $rew'(\langle s, r \rangle, \alpha) = 0$ .
  - If  $r' > R$  then  $P'(\langle s, r \rangle, \alpha, t) = P(s, \alpha, t)$  for all states  $t \in S$  and  $rew'(\langle s, r \rangle, \alpha) = 0$ .
- If the current state of  $\mathcal{M}'$  is a state  $\langle goal, r \rangle$  then  $\mathcal{M}'$  switches to the goal state in the second mode while earning reward  $r$ . That is,  $P'(\langle goal, r \rangle, \tau, goal) = 1$  and  $rew'(\langle goal, r \rangle, \tau) = r$  for some distinguished action name  $\tau \in Act$ .
- If the current state in the new MDP  $\mathcal{M}'$  is a state  $s \in S$  then  $\mathcal{M}'$  behaves as  $\mathcal{M}$ , i.e.,  $Act_{\mathcal{M}'}(s) = Act_{\mathcal{M}}(s)$ ,  $P'(s, \alpha, t) = P(s, \alpha, t)$  and  $rew'(s, \alpha) = rew(s, \alpha)$  if  $s, t \in S$  and  $\alpha \in Act$ .

The states  $goal$  and  $fail$  and  $\langle fail, r \rangle$  for  $r \in \{0, 1, \dots, R\}$  are trap states.

Obviously, there is a one-to-one correspondence between the paths in  $\mathcal{M}$  and the paths in  $\mathcal{M}'$  and between the schedulers of  $\mathcal{M}$  and  $\mathcal{M}'$ . Given a finite path  $\pi'$  in  $\mathcal{M}'$ , let  $\pi'|_{\mathcal{M}}$  denote the path in  $\mathcal{M}$  resulting from  $\pi'$  by dropping the annotations for the first mode. Vice versa, for a given path  $\pi$  in  $\mathcal{M}$  we add appropriate annotations for the states in some prefix of  $\pi$  to obtain a path  $lift(\pi)$  in  $\mathcal{M}'$  starting in  $s'_{init} = \langle s_{init}, 0 \rangle$  with  $lift(\pi)|_{\mathcal{M}} = \pi$ . We have:

- If  $\pi'$  consists of states in the first mode, i.e., all states of  $\pi'$  are annotated states in  $S \times \{0, 1, \dots, R\}$ , then  $rew'(\pi') = 0$  and  $rew(\pi'|_{\mathcal{M}}) = r$  where  $last(\pi') = \langle s, r \rangle$ .
- If  $last(\pi') \in S$  is a state in second mode then  $rew'(\pi') = rew(\pi'|_{\mathcal{M}})$ .
- If the last transition in  $\pi'$  is a mode-switch, say  $\langle s, r \rangle \xrightarrow{\alpha} t$ , then  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$  for each scheduler  $\mathfrak{S}$  of  $\mathcal{M}$  where  $\pi'|_{\mathcal{M}}$  is a  $\mathfrak{S}$ -path (otherwise  $\mathfrak{S}$  would be a critical scheduler).

Obviously,  $\mathcal{M}$  and  $\mathcal{M}'$  have the same conditional expected accumulated reward to reach  $goal$  under corresponding schedulers. In particular:

$$\mathbb{CE}^{\max} = \mathbb{E}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal \mid \Diamond goal) = \mathbb{E}_{\mathcal{M}', \langle s_{init}, 0 \rangle}^{\max}(\Diamond goal \mid \Diamond goal)$$

The MDP  $\mathcal{N}$  results from  $\mathcal{M}'$  by adding reset-transitions from the states  $fail$  and  $\langle fail, \mathcal{G} \rangle$ . Let

$$Fail = \{ fail \} \cup \{ \langle fail, r \rangle : r \in \{0, 1, \dots, R\} \}$$

For all states  $s_f \in \text{Fail}$  we deal with  $\text{Act}_{\mathcal{N}}(s_f) = \{\tau\}$  for some distinguished action  $\tau$  and define  $P_{\mathcal{N}}(s_f, \tau, s'_{\text{init}}) = 1$  and  $\text{rew}_{\mathcal{N}}(s_f, \tau) = 0$ . The transition probabilities of all other states and the rewards of all other state-action pairs are the same as in  $\mathcal{M}'$ .

While  $\mathcal{M}'$  has no end components (by assumption (A2)),  $\mathcal{N}$  might have end components containing  $s'_{\text{init}}$  and at least one of the fail-states. The above shows that the reward of each finite path in  $\mathcal{M}'$  from  $s'_{\text{init}}$  to some state  $\langle \text{fail}, r \rangle$  is 0, while the reward of paths from  $s'_{\text{init}}$  to  $\text{fail}$  can be positive. We now show that the end components of  $\mathcal{N}$  cannot contain any of the states of  $\mathcal{M}$ 's second mode. In particular, there is no end component containing state  $\text{fail}$ .

*Claim.*  $\mathcal{N}$  has no positive end components.

*Proof of the claim.* By the definition of the reward function in  $\mathcal{N}$ , it suffices to show that none of the states  $t \in S \setminus \{\text{goal}\}$  belongs to an end component of  $\mathcal{N}$ .

Suppose by contradiction that there is an end component  $\mathcal{E} = (E, \mathfrak{A})$  of  $\mathcal{N}$  containing some state  $t \in S \setminus \{\text{goal}\}$ . Since  $\mathcal{M}$  and  $\mathcal{M}'$  do not contain end components (assumption (A2)),  $\mathcal{E}$  must contain one of the reset-transitions from  $\text{fail}$  or some state  $\langle \text{fail}, r \rangle$  to the initial state  $s'_{\text{init}} = \langle s_{\text{init}}, 0 \rangle$ . As  $\text{goal}$  is a trap,  $\mathcal{E}$  does not contain  $\text{goal}$  and none of its copies  $\langle \text{goal}, r' \rangle$ . We pick a finite-memory scheduler  $\mathfrak{S}'$  for  $\mathcal{N}$  such that

$$\varsigma \models \Box \Diamond t \text{ and } \varsigma \models \Box E \text{ for all infinite } \mathfrak{S}'\text{-paths } \varsigma \text{ starting in } s'_{\text{init}}.$$

Thus,  $\Pr_{\mathcal{N}, s'_{\text{init}}}^{\mathfrak{S}'}(\Box E) = 1$  and therefore  $\Pr_{\mathcal{N}, s'_{\text{init}}}^{\mathfrak{S}'}(\Diamond \text{goal}) = 0$ .

Let  $\pi'$  be a shortest  $\mathfrak{S}'$ -path from  $s'_{\text{init}}$  to  $t$ . Clearly,  $\pi'$  contains a mode-switch. Hence, it must contain a positive cycle. As  $\mathcal{M}$  does not have critical schedulers, the scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  that behaves as  $\mathfrak{S}'$  as long as none of the fail-states has been visited enjoys the property  $\Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{S}}(\Diamond \text{goal}) > 0$ . As all  $\mathfrak{S}$ -paths are  $\mathfrak{S}'$ -paths we get

$$\Pr_{\mathcal{N}, s'_{\text{init}}}^{\mathfrak{S}'}(\Diamond \text{goal}) > 0$$

Contradiction. This completes the proof of the claim.

Using the results of [11], each scheduler  $\mathfrak{S}'$  for  $\mathcal{M}'$  induces a corresponding scheduler  $\mathfrak{S}$  for  $\mathcal{N}$  such that the conditional probability for some reachability objective  $\varphi$  under the assumption  $\Diamond \text{goal}$  with respect to scheduler  $\mathfrak{S}'$  in  $\mathcal{M}'$  agrees with the unconditional probability for  $\varphi$  in  $\mathcal{N}$  with respect to scheduler  $\mathfrak{S}$ . If  $\pi'$  is a  $\mathfrak{S}'$ -path from  $s'_{\text{init}} = \langle s_{\text{init}}, 0 \rangle$  to  $\text{goal}$  then  $\pi'$  is a suffix of all “corresponding”  $\mathfrak{S}$ -paths  $\pi$  in  $\mathcal{N}$  and therefore  $\text{rew}'(\pi') \leq \text{rew}_{\mathcal{N}}(\pi)$ . With Lemma C.5 applied to  $\mathcal{N}$  we get:

$$\mathbb{E}_{\mathcal{M}', s'_{\text{init}}}^{\max}(\Diamond \text{goal} \mid \Diamond \text{goal}) \leq \mathbb{E}_{\mathcal{N}, \langle s_{\text{init}}, 0 \rangle}^{\max}(\Diamond \text{goal}) < \infty$$

This completes the proof of Proposition C.8. ■

**Corollary C.9.** *If  $\mathcal{M}$  is an MDP where (A1) and (A2) hold and  $\Pr_s^{\min}(\Diamond \text{goal}) > 0$  for all states  $s \in S \setminus \{\text{fail}\}$  then  $\mathbb{CE}^{\max} < \infty$ .*

*Proof.* The proof is obvious as there is no scheduler  $\mathfrak{U}$  with  $\Pr_{s_{\text{init}}}^{\mathfrak{U}}(\Box \neg \text{goal}) = 1$ . Hence, there is no critical scheduler. ■

### C.3 Algorithm for checking finiteness of $\mathbb{CE}^{\max}$

By the obtained results, the finiteness of the maximal conditional expected accumulated reward can be checked as follows. We first apply standard algorithms to compute the maximal end components. If one of them is positive then the maximal conditional expected accumulated reward is infinite (see Lemma C.4). Otherwise we switch from  $\mathcal{M}$  to its MEC-quotient, i.e., we collapse all maximal end components into a single state and identify  $\mathcal{M}$  with the resulting MDP. The remaining task is to check the absence of critical schedulers in  $\mathcal{M}$  (see Definition C.7 and Proposition C.8). For this, we consider the largest sub-MDP  $\mathcal{M}_{\setminus goal}$  that results by iteratively removing states and actions. We first remove state *goal* and for each state  $s$ , we remove all actions  $\alpha$  from  $Act(s)$  with  $P(s, \alpha, goal) > 0$ . If some state arises where  $Act(s)$  is empty then we remove state  $s$  using the same technique. For all states  $s$  in the resulting sub-MDP  $\mathcal{M}_{\setminus goal}$  we have  $\Pr_{\mathcal{M}_{\setminus goal}, s}^{\min}(\Diamond fail) = 1$  (by (A2)) and  $\Pr_{\mathcal{M}, s}^{\max}(\Diamond fail) = 1$  (as  $\Pr_{\mathcal{M}, s}^{\mathfrak{S}}(\Diamond fail) = 1$  for each scheduler for  $\mathcal{M}_{\setminus goal}$  viewed as a scheduler of  $\mathcal{M}$  with initial state  $s$ ). Furthermore,  $\mathcal{M}_{\setminus goal}$  has a positive cycle if and only if  $\mathcal{M}$  has a critical scheduler. The existence of a positive cycle can be checked using a nested depth-first search [22] as it is known for checking the existence of a path satisfying a Büchi (repeated reachability) condition in ordinary transition systems. Together with the preprocessing explained in Appendix C.1 we get:

**Corollary C.10.** *The task to check whether  $\mathbb{CE}^{\max}$  is finite is solvable in time polynomial in the size of  $\mathcal{M}$ .*

### C.4 Computing an upper bound

Suppose now that  $\mathcal{M}$  has no critical schedulers, in which case  $\mathbb{CE}^{\max}$  is finite. The maximal unconditional expected accumulated reward until goal in the MDP  $\mathcal{N}$  constructed in the proof of Proposition C.8 yields an upper bound  $\mathbb{CE}^{\text{ub}}$  for the maximal conditional expected accumulated reward until reach *goal* in  $\mathcal{M}$ :

$$\mathbb{CE}^{\text{ub}} \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{N}, \langle s_{\text{init}}, 0 \rangle}^{\max}(\Diamond goal) \geq \mathbb{CE}^{\max}$$

We now address the complexity bounds for the computation of  $\mathbb{CE}^{\text{ub}}$  stated in Theorem 1. The logarithmic length of  $R = \sum_s rew^{\max}(s)$  where  $rew^{\max}(s) = \max\{rew(s, \alpha) : \alpha \in Act(s)\}$  is linear in  $size(\mathcal{M})$ . The size of the MDPs  $\mathcal{N}$  is in  $\mathcal{O}(R \cdot size(\mathcal{M}))$ . The computation of the upper bound  $\mathbb{CE}^{\text{ub}}$  then corresponds to the computation of the maximal unconditional expected accumulated reward to reach *goal* in  $\mathcal{N}$ , which has a polynomial time bound in the size of  $\mathcal{N}$ . Consequently,  $\mathbb{CE}^{\text{ub}}$  can be computed in time polynomial in  $R$  and the size of  $\mathcal{M}$ , which gives the pseudo-polynomial time bound as stated in Theorem 1.

As noted in the proof of Prop. C.8, for the special case where  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond goal) > 0$  for all states  $s \in S \setminus \{fail\}$ , we can avoid the annotation of the states with the accumulated reward up to  $N$ . In this case, the size of  $\mathcal{N}$  is polynomial in  $size(\mathcal{M})$ , which leads to the polynomial bound on the computation time of  $\mathbb{CE}^{\text{ub}}$  as stated in Theorem 1.

## D Deterministic reward-based schedulers are sufficient

In this section, we show that deterministic reward-based schedulers are sufficient for  $\mathbb{CE}^{\max}$  where we suppose that the given MDP  $\mathcal{M} = (S, Act, P, s_{init}, rew)$  has two trap states *goal* and *fail* and satisfies (A1), (A2) and  $\mathbb{CE}^{\max} < \infty$ . Recall that (A1) asserts that *goal* is reachable from all states  $s \in S \setminus \{fail\}$ , while (A2) asserts that  $\mathcal{M}$  has no end components and therefore  $\Pr_{\mathcal{M},s}^{\min}(\Diamond(goal \vee fail)) = 1$  for all states  $s \in S$ .

Recall that deterministic reward-based schedulers can be viewed as functions  $\mathfrak{S} : S \times \mathbb{N} \rightarrow Act$ .

### Proposition D.1.

$$\mathbb{CE}^{\max} = \sup \left\{ \mathbb{CE}^{\mathfrak{S}} : \mathfrak{S} \text{ is a deterministic reward-based scheduler for } \mathcal{M} \text{ such that } \Pr_{\mathcal{M},s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0 \right\}$$

*Proof.* We consider the MDPs  $\mathcal{M}'$  and  $\mathcal{N}$  that have been introduced in the proof of Proposition C.8. Recall that  $\mathcal{M}$  and  $\mathcal{M}'$  have the same maximal conditional expectation and there is a one-to-one-correspondence between the reward-based schedulers for  $\mathcal{M}$  and  $\mathcal{M}'$ .<sup>8</sup> For the following arguments we may identify  $\mathcal{M}$  and  $\mathcal{M}'$ . This means that we may assume whenever  $\mathfrak{S}$  is a scheduler for  $\mathcal{M}$  with  $\Pr_{\mathcal{M},s_{init}}^{\mathfrak{S}}(\Diamond fail) = 1$  then  $rew(\pi) = 0$  for all  $\mathfrak{S}$ -paths from  $s_{init}$  to *fail*. Stated differently, by identifying  $\mathcal{M}$  and  $\mathcal{M}'$  we have  $\Pr_{\mathcal{M},s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$  for each scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  where  $rew(\pi) > 0$  for some  $\mathfrak{S}$ -path from  $s_{init}$  to *fail*. With these assumptions, the MDP  $\mathcal{N}$  arises from  $\mathcal{M}$  by adding a reset transition from *fail* to  $s_{init}$  with reward 0. As shown in the proof of Proposition C.8,  $\mathcal{N}$  has no positive end components, and therefore  $\mathbb{E}_{\mathcal{N},s_{init}}^{\max}(\Diamond goal)$  is finite. Moreover, as *goal* is the only trap state in  $\mathcal{N}$  and as all state-action pairs contained in some end component of  $\mathcal{N}$  have reward 0,  $\mathbb{E}_{\mathcal{N},s_{init}}^{\max}(\Diamond goal)$  agrees with the maximal total reward in  $\mathcal{N}$ .

We now switch from  $\mathcal{N}$  to a new MDP  $\mathcal{N}_{acc}$  that arises from  $\mathcal{N}$  by attaching the accumulated reward to all states and modifying  $\mathcal{N}$ 's reward structure such that the reward of the reset transition after a path  $\pi$  from  $s_{init}$  to a fail state is  $-rew(\pi)$ . Thus,  $\mathcal{N}_{acc}$  has infinitely many states and positive and negative rewards. Formally, the definition of  $\mathcal{N}_{acc}$  is as follows. The state space of  $\mathcal{N}_{acc}$  is  $S_{acc} = S \times \mathbb{N}$ . The action set remains unchanged, i.e.,  $Act_{\mathcal{N}_{acc}} = Act_{\mathcal{N}} = Act \cup \{\tau\}$  with  $\tau$  being a symbol for the reset transitions. The transition probabilities and reward structure in  $\mathcal{N}_{acc}$  are given by:

$$P_{acc}(\langle s, r \rangle, \alpha, \langle s', r' \rangle) = \begin{cases} P(s, \alpha, s') & : \text{if } r' = r + rew(s, \alpha) \\ 0 & : \text{otherwise} \end{cases}$$

<sup>8</sup> If a deterministic reward-based scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  chooses action  $\alpha$  for  $(s, r)$  with  $r \leq R$  then the lifted scheduler for  $\mathcal{M}'$  chooses  $\alpha$  for the state-action pair  $(\langle s, r \rangle, 0)$  in  $\mathcal{M}'$ , and vice versa. Recall that the accumulated reward of each path in  $\mathcal{M}'$  from  $s'_{init} = \langle s_{init}, 0 \rangle$  to some state  $\langle s, r \rangle$  in the first mode has reward 0. For the state-reward pairs  $\langle s', r \rangle$  where  $s'$  is a state of  $\mathcal{M}'$  in its second mode (i.e.,  $s' \in S$ ), the lifted scheduler selects the same action as  $\mathfrak{S}$ .

and

$$rew_{acc}(\langle s, r \rangle, \alpha) = rew(s, \alpha)$$

for all states  $s \in S \setminus \{goal, fail\}$ ,  $s' \in S$ , actions  $\alpha \in Act(s)$  and  $r, r' \in \mathbb{N}$ . The reset action  $\tau$  is the only enabled action of the fail states with

$$P_{acc}(\langle fail, r \rangle, \tau, \langle s_{init}, 0 \rangle) = 1 \quad \text{and} \quad rew_{acc}(\langle fail, r \rangle, \tau) = -r$$

and  $P_{acc}(\cdot) = 0$  in all remaining cases. The starting state of  $\mathcal{N}_{acc}$  is  $\langle s_{init}, 0 \rangle$ .

Each path in the original MDP  $\mathcal{M}$  can be lifted to a path  $lift(\pi)$  in  $\mathcal{N}_{acc}$  by augmenting the states with reward values. Vice versa, for each finite path  $\pi' = \langle s_0, 0 \rangle \alpha_0 \langle s_1, r_1 \rangle \alpha_1 \dots \alpha_{n-1} \langle s_n, r_n \rangle$  in  $\mathcal{N}_{acc}$  where  $fail \notin \{s_0, s_1, \dots, s_n\}$ , the sequence  $\pi = s_0 \alpha_0 s_1 \alpha_1 \dots \alpha_{n-1} s_n$  is a path in  $\mathcal{M}$  with  $rew(\pi) = r_n$ . Thus,  $rew_{acc}(\pi') = 0$  for all paths  $\pi'$  in  $\mathcal{N}_{acc}$  from  $\langle s_{init}, 0 \rangle$  to  $\langle s_{init}, 0 \rangle$  where the last transition is a reset-transition from some fail state  $\langle fail, r \rangle$  to  $\langle s_{init}, 0 \rangle$ . If now  $\pi'$  is an arbitrary path in  $\mathcal{N}_{acc}$  from  $\langle s_{init}, 0 \rangle$  to some state  $\langle s, r \rangle$  then  $rew_{acc}(\pi') = r$ . We define  $\pi'|_{\mathcal{M}}$  as the unique path  $\pi$  from  $s_{init}$  to  $s$  in  $\mathcal{M}$  that arises from  $\pi'$  by (i) removing the longest prefix  $\varrho$  of  $\pi'$  where the last transition of  $\varrho$  is a reset transition (i.e.,  $\pi'$  has the form  $\varrho; \pi''$  where  $\pi''$  is a path from  $\langle s_{init}, 0 \rangle$  to some state  $\langle s, r \rangle$  in  $\mathcal{N}_{acc}$  that does not contain a reset transition) and (ii) erasing the reward annotations of remaining path  $\pi''$ . Then,  $rew_{acc}(\pi') = rew(\pi'|_{\mathcal{M}})$  and the lifting of  $\pi'|_{\mathcal{M}}$  is the suffix  $\pi''$  of  $\pi'$ .

Each deterministic memoryless scheduler  $\mathfrak{S}$  for  $\mathcal{N}_{acc}$  can be viewed as a deterministic reward-based scheduler for  $\mathcal{M}$ . That is, if  $\mathfrak{S}'$  is a deterministic memoryless scheduler for  $\mathcal{N}_{acc}$  then the corresponding deterministic reward-based scheduler  $\mathfrak{S}'|_{\mathcal{M}}$  for  $\mathcal{M}$  is given by  $\mathfrak{S}'|_{\mathcal{M}}(s, r) = \mathfrak{S}'(\langle s, r \rangle)$ . Using arguments as in [11] (see also Section B), we get:

$$\mathbb{E}_{\mathcal{N}_{acc}, s_{init}}^{\mathfrak{S}'}(\Diamond goal) = \mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}'|_{\mathcal{M}}}(\Diamond goal \mid \Diamond goal)$$

Vice versa, each (possibly history-dependent and randomized) scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  induces a scheduler  $lift(\mathfrak{S})$  for  $\mathcal{N}_{acc}$  given by  $lift(\mathfrak{S})(\pi') = \mathfrak{S}(\pi'|_{\mathcal{M}})$ . Thus, the residuals of the lifted scheduler enjoy the property  $lift(\mathfrak{S})\uparrow\pi' = lift(\mathfrak{S})$  for each finite path  $\pi'$  in  $\mathcal{N}_{acc}$  where the last transition in  $\pi'$  is the reset transition from some fail state  $\langle fail, r \rangle$  to  $\langle s_{init}, 0 \rangle$ . Again, using [11], we obtain:

$$\mathbb{E}_{\mathcal{N}_{acc}, s_{init}}^{lift(\mathfrak{S})}(\Diamond goal) = \mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal \mid \Diamond goal)$$

The remaining task is to show that the supremum of the values  $\mathbb{E}_{\mathcal{N}_{acc}, s_{init}}^{\mathfrak{S}'}(\Diamond goal)$  when ranging over all schedulers  $\mathfrak{S}'$  for  $\mathcal{N}_{acc}$  agrees with the supremum over the deterministic memoryless schedulers for  $\mathcal{N}_{acc}$ . For this, we rely on known results for infinite-state MDPs with positive and negative rewards as stated e.g. in [33]. Let  $\mathcal{N}_{acc}^+$  be the MDP resulting from  $\mathcal{N}_{acc}$  by replacing the reward function  $rew_{acc}$  with the (non-negative) reward function  $rew_{acc}^+$  defined by

$$rew_{acc}^+(\langle s, r \rangle, \alpha) = \max \{ rew_{acc}(s, \alpha), 0 \} = rew_{\mathcal{N}}(s, \alpha)$$

for all states  $s \in S \setminus \{goal\}$ , actions  $\alpha \in Act_{\mathcal{N}}(s)$  and  $r \in \mathbb{N}$ . In particular,  $rew_{acc}^+(\langle fail, r \rangle, \tau) = 0$ . It is well-known (see Section 7.1 in [33]) that if the maximal (unconditional) expected total reward in  $\mathcal{N}_{acc}^+$  is finite, then the supremum of the expected total reward in  $\mathcal{N}_{acc}$  when ranging over all schedulers  $\mathfrak{S}'$  for  $\mathcal{N}_{acc}$  agrees with the supremum over the deterministic memoryless schedulers for  $\mathcal{N}_{acc}$ . The expected total reward in  $\mathcal{N}_{acc}$  under some scheduler  $\mathfrak{S}'$  agrees with the value  $\mathbb{E}_{\mathcal{N}_{acc}, \langle s_{init}, 0 \rangle}^{\mathfrak{S}' }(\Diamond Goal)$  where  $Goal = \{\langle goal, r \rangle : r \in \mathbb{N}\}$ . The analogous statement holds for  $\mathcal{N}_{acc}^+$ . So, the remaining task is to show that the maximal expected reward until reaching a goal state in  $\mathcal{N}_{acc}^+$  is finite. The relation  $\mathcal{R} = \{(s, \langle s, r \rangle) : r \in \mathbb{N}\}$  is a reward-preserving bisimulation for the MDPs  $\mathcal{N}_{acc}^+$  and the MDP  $\mathcal{N}$  resulting from  $\mathcal{M}$  by adding a reset-transition of reward 0 from  $fail$  to  $s_{init}$  (see above). Hence:

$$\mathbb{E}_{\mathcal{N}_{acc}^+, \langle s_{init}, 0 \rangle}^{\max}(\Diamond Goal) = \mathbb{E}_{\mathcal{N}, s_{init}}^{\max}(\Diamond goal) < \infty$$

As stated above, using classical results about countable MDPs with positive and negative rewards (see e.g. [33]), we obtain:

$$\begin{aligned} & \mathbb{E}_{\mathcal{N}_{acc}, s_{init}}^{\max}(\Diamond Goal) \\ &= \sup \{ \mathbb{E}_{\mathcal{N}_{acc}, \langle s_{init}, 0 \rangle}^{\mathfrak{S}'}(\Diamond Goal) : \mathfrak{S}' \text{ is a det. memoryless scheduler for } \mathcal{N}_{acc} \} \\ &= \sup \{ \mathbb{E}_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal) : \mathfrak{S} \text{ is a det. reward-based scheduler for } \mathcal{M} \} \end{aligned}$$

This completes the proof of Proposition D.1.  $\blacksquare$

## E Existence of a saturation point and optimal schedulers

The main obligation in this section is to provide a proof for Proposition 4.1 in the main paper, asserting the existence of an optimal reward-based scheduler that is memoryless for sufficiently large accumulated rewards.

In this and the remaining sections of the appendix, we suppose that MDP  $\mathcal{M} = (S, Act, P, s_{init}, rew)$  has two trap states  $goal$  and  $fail$  and satisfies assumptions (A1), (A2) stated in Appendix C.2 and  $\mathbb{CE}^{\max} < \infty$ . Recall that (A1) asserts that  $goal$  is reachable from all states  $s \in S \setminus \{fail\}$ , while (A2) asserts that  $\mathcal{M}$  has no end components and therefore  $\Pr_{\mathcal{M}, s}^{\min}(\Diamond(goal \vee fail)) = 1$  for all states  $s \in S$ .

In what follows, we will often use the residual notations  $\mathfrak{S} \uparrow R$ ,  $\mathfrak{S} \uparrow(s, R)$  and the redefine operator  $\mathfrak{S} \triangleleft_R \mathfrak{T}$  that have been introduced for reward-based schedulers. See Section A.

**Proposition E.1 (See Proposition 4.1).** *There exists a natural number  $\wp$  (called saturation point of  $\mathcal{M}$ ) and a deterministic memoryless scheduler  $\mathfrak{M}$  such that:*

$$(a) \quad \mathbb{CE}^{\mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}} \text{ for each scheduler } \mathfrak{T} \text{ with } \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{T}}(\Diamond goal) > 0.$$

(b)  $\mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\max}$  for some deterministic reward-based scheduler  $\mathfrak{S}$  such that  $\Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{S}}(\Diamond \text{goal}) > 0$  and  $\mathfrak{S} \upharpoonright_{\wp} = \mathfrak{M}$ .

In this section the statement (a) is captured in Lemma E.16, whereas statement (b) corresponds to Lemma E.15. In order to prove Prop. E.1, we first show the existence of a saturation point  $\wp$  which will be derived – among others – from the convergence rate for reaching one of the trap states (see Lemma E.4 below). The so obtained saturation point is, however, very large. Later (see Section F), we show that there is a smaller and easily computable saturation point as outlined in Section 4.

### E.1 Some technical statements

The following lemmas are trivial observations about quotients of sums that will be used at various places for comparing the conditional expectations under different schedulers.

**Lemma E.2.** *For all real numbers  $K, L, k, l$  with  $L > 0$  and  $l > 0$ , one of the following three cases applies:*

$$\begin{aligned} & \frac{K}{L} < \frac{K+k}{L+l} < \frac{k}{l} \\ \text{or} \quad & \frac{K}{L} > \frac{K+k}{L+l} > \frac{k}{l} \\ \text{or} \quad & \frac{K}{L} = \frac{K+k}{L+l} = \frac{k}{l} \end{aligned}$$

The proof of Lemma E.2 is straightforward and omitted here. Note, however, if  $K/L < k/l \leq k'/l'$  then  $(K+k')/(L+l') < (K+k)/(L+l)$  is possible. Consider, for example,  $K = 1, L = 2$  and  $k = l = 2$  and  $k' = l' = 1$ . Then,  $(K+k')/(L+l') = 2/3 < 3/4 = (K+k)/(L+l)$ .

**Lemma E.3.** *Let  $\rho, \theta, \zeta, x, y, z$  be real numbers such that  $x, y$  and  $z$  are non-negative and  $x + y > 0, x + z > 0$  and  $y > z$ . Then, one of the following three cases holds:*

$$\begin{aligned} & \frac{\rho + \zeta}{x + z} < \frac{\rho + \theta}{x + y} < \frac{\theta - \zeta}{y - z} \\ \text{or} \quad & \frac{\rho + \zeta}{x + z} > \frac{\rho + \theta}{x + y} > \frac{\theta - \zeta}{y - z} \\ \text{or} \quad & \frac{\rho + \zeta}{x + z} = \frac{\rho + \theta}{x + y} = \frac{\theta - \zeta}{y - z} \end{aligned}$$

*Proof.* The claim follows from Lemma E.2 with  $(K, L) = (\rho + \zeta, x + z)$  and  $(k, l) = (\theta - \zeta, y - z)$ .  $\blacksquare$

We often make use of Lemma E.3 in the following form. If  $y > z$  then

$$\frac{\rho + \theta}{x + y} \leq \frac{\rho + \zeta}{x + z} \quad \text{iff} \quad \frac{\theta - \zeta}{y - z} \leq \frac{\rho + \theta}{x + y} \quad \text{iff} \quad \frac{\theta - \zeta}{y - z} \leq \frac{\rho + \zeta}{x + z}$$

and the analogous statements for other comparison operators  $<$ ,  $>$ ,  $\geq$  and  $=$  rather than  $\leq$ .

## E.2 Convergence rate

**Lemma E.4 (Fast convergence for reaching a trap).** *There exists  $\lambda \in ]0, 1[$  and  $R_0 \in \mathbb{N}$  such that for each state  $s \in S \setminus \{goal, fail\}$  and  $r \in \mathbb{N}$  with  $r \geq R_0$ :*

$$\Pr_{\mathcal{M},s}^{\min} ( \Diamond^{\leq r} (goal \vee fail) ) \geq 1 - \lambda^r$$

*Proof.* The proof relies on a calculation similar to the one of [36]. The parameter  $\lambda$  depends on the minimal positive transition probability, the maximal reward assigned to a state-action pair and the number of states in  $\mathcal{M}$ . Let  $N = |S|$  and

$$q = \min \{ P(s, \alpha, t) : (s, \alpha, t) \in S \times Act \times S, P(s, \alpha, t) > 0 \}$$

$$R = \max \{ rew(s, \alpha) : (s, \alpha) \in S \times Act \}$$

(A2) yields that the probability to reach a trap state *goal* or *fail* from any state  $s$  within  $N$  or fewer steps is at least  $p^N$  under each scheduler. The accumulated reward  $rew(\pi)$  of paths  $\pi$  with  $|\pi| \leq N$  is bounded by  $N \cdot R$ . Thus, for all schedulers  $\mathfrak{S}$  and all  $r, k \in \mathbb{N}$  with  $r = k \cdot N \cdot R$ :

$$\Pr_s^{\mathfrak{S}} ( \neg(\Diamond^{\leq r} (goal \vee fail)) ) \leq (1 - q^N)^k$$

Let  $R_0 = NR$  and  $\lambda = (1 - q^N)^{1/2NR}$  if  $q < 1$ . Then,  $0 < \lambda < 1$  and for all  $r \geq R_0$ :

$$\lambda^r = (1 - q^N)^{r/2NR} \geq (1 - q^N)^{\lfloor r/NR \rfloor}$$

Hence, for each scheduler  $\mathfrak{S}$  and  $r \geq R_0$ :

$$\begin{aligned} \Pr_{\mathcal{M},s}^{\mathfrak{S}} ( \Diamond^{\leq r} (goal \vee fail) ) &\geq \Pr_{\mathcal{M},s}^{\mathfrak{S}} ( \Diamond^{\leq \lfloor r/NR \rfloor NR} (goal \vee fail) ) \\ &= 1 - \Pr_{\mathcal{M},s}^{\mathfrak{S}} ( \neg \Diamond^{\leq \lfloor r/NR \rfloor NR} (goal \vee fail) ) \\ &\geq 1 - (1 - q^N)^{\lfloor r/NR \rfloor} \geq 1 - \lambda^r \end{aligned}$$

If  $q = 1$  then MDP can be viewed as a nondeterministic transition system, in which case we can deal with any  $\lambda \in ]0, 1[$ .  $\blacksquare$

As Remark C.3 shows, assumption (A2) is crucial for Lemma E.4. As a consequence of Lemma E.4 we get that if  $\mathcal{M}$  satisfies assumptions (A1) and (A2) then  $\mathbb{CE}^{\mathfrak{S}} < \infty$  for each scheduler  $\mathfrak{S}$  where  $\Pr_{s_{init}}^{\mathfrak{S}} (\Diamond goal)$  is positive, but the supremum over all schedulers can still be infinite (see Proposition C.8).



**Lemma E.5 (Arrearage of expected accumulated rewards).** *Let  $\lambda$  and  $R_0$  be as in Lemma E.4. Then, for each  $R \geq R_0$ , each scheduler  $\mathfrak{S}$  and each state  $s$  of  $\mathcal{M}$  we have:*

$$\sum_{r=R}^{\infty} r \cdot \Pr_{\mathcal{M},s}^{\mathfrak{S}}(\diamond^{=r} goal) \leq C \cdot R \cdot \lambda^R \quad \text{where} \quad C = \frac{1}{(1-\lambda)^2}$$

*Proof.* As  $0 < \lambda < 1$ , the infinite series  $\sum_{r=0}^{\infty} r \cdot \lambda^r$  converges to  $\lambda/(1-\lambda)^2$ . More precisely, for each  $R \in \mathbb{N}$  with  $R \geq R_0$ :

$$\begin{aligned} \sum_{r=R}^{\infty} r \cdot \lambda^r &= \lambda^R \cdot \sum_{r=0}^{\infty} (r+R) \cdot \lambda^r = \lambda^R \cdot \sum_{r=0}^{\infty} r \cdot \lambda^r + R \cdot \lambda^R \cdot \sum_{r=0}^{\infty} \lambda^r \\ &= \frac{\lambda^{R+1}}{(1-\lambda)^2} + \frac{R \cdot \lambda^R}{1-\lambda} \\ &= R \cdot \lambda^R \cdot \left( \frac{1}{R} \cdot \frac{\lambda}{(1-\lambda)^2} + \frac{1}{1-\lambda} \right) \\ &\leq R \cdot \lambda^R \cdot \left( \frac{\lambda}{(1-\lambda)^2} + \frac{1}{1-\lambda} \right) = R \cdot \lambda^R \cdot \frac{1}{(1-\lambda)^2} = C \cdot R \cdot \lambda^R \end{aligned}$$

This yields the claim.  $\blacksquare$

Recall (see Definition C.2) that if  $s \in S \setminus \{goal, fail\}$  and  $\mathfrak{T}$  is a scheduler with  $\Pr_s^{\mathfrak{T}}(\diamond goal) > 0$  then:

$$\mathbb{CE}_s^{\mathfrak{T}} = \mathbb{E}_s^{\mathfrak{T}}(\diamond goal \mid \diamond goal) = \frac{\mathbb{E}_s^{\mathfrak{T}}}{\Pr_s^{\mathfrak{T}}(\diamond goal)}$$

where  $\mathbb{E}_s^{\mathfrak{T}}$  is a shorthand notation for  $\mathbb{E}_s^{\mathfrak{T}}(\diamond goal)$  given by

$$\mathbb{E}_s^{\mathfrak{T}} = \sum_{r=0}^{\infty} r \cdot \Pr_s^{\mathfrak{T}}(\diamond^{=r} goal)$$

**Corollary E.6.** *Assumptions and notations as in Lemma E.5. For each scheduler  $\mathfrak{T}$ , each  $R \in \mathbb{N}$  with  $R \geq R_0$  and each state  $s \in S \setminus \{goal, fail\}$  we have:*

$$\mathbb{E}_s^{\mathfrak{T}} \leq (R-1) \cdot \Pr_s^{\mathfrak{T}}(\diamond^{<R} goal) + C \cdot R \cdot \lambda^R$$

Moreover, if  $\Pr_s^{\mathfrak{T}}(\diamond goal) > 0$  and  $\mathbb{CE}_s^{\mathfrak{T}} \geq 2R$  then  $\Pr_s^{\mathfrak{T}}(\diamond goal) < C \cdot \lambda^R$ .

*Proof.* Clearly, for each scheduler  $\mathfrak{T}$  and state  $s$  we have:

$$\sum_{r=0}^{R-1} r \cdot \Pr_s^{\mathfrak{T}}(\diamond^{=r} goal) \leq (R-1) \cdot \Pr_s^{\mathfrak{T}}(\diamond^{<R} goal)$$

Thus, the first statement is a consequence of Lemma E.5. For the second statement, we suppose  $\mathbb{CE}_s^{\mathfrak{T}} \geq 2R$ . But then:

$$2R \cdot \Pr_s^{\mathfrak{T}}(\diamond goal) \leq \mathbb{E}_s^{\mathfrak{T}} < R \cdot \Pr_s^{\mathfrak{T}}(\diamond goal) + C \cdot R \cdot \lambda^R$$

and therefore:  $R \cdot \Pr_s^{\mathfrak{T}}(\diamond goal) < C \cdot R \cdot \lambda^R$ . This yields  $\Pr_s^{\mathfrak{T}}(\diamond goal) < C \cdot \lambda^R$ .  $\blacksquare$

Recall that we use the notation  $\Diamond^{\bowtie n}$  for reward-bounded reachability, while  $\bigcirc^{\bowtie n}$  denotes a step-bound.

**Proposition E.7 (Continuity of conditional expectations).** *Let  $\mathfrak{S}$  be a scheduler of  $\mathcal{M}$  with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$  and  $\varepsilon > 0$ . Then, there exists  $n_\varepsilon \in \mathbb{N}$  such that  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\bigcirc^{\leq n_\varepsilon} goal) > 0$  and for each scheduler  $\mathfrak{T}$  with  $\mathfrak{S}(\pi) = \mathfrak{T}(\pi)$  for all finite paths  $\pi$  with  $|\pi| \leq n_\varepsilon$  and  $first(\pi) = s_{init}$  we have:*

$$|\mathbb{CE}^{\mathfrak{S}} - \mathbb{CE}^{\mathfrak{T}}| < \varepsilon$$

Note that under the above assumption  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{T}}(\Diamond goal)$  is positive.

*Proof.* Let  $\Pi_n$  denote the set of all finite paths  $s_0 \alpha_0 s_1 \dots \alpha_{n-1} s_n$  of length  $n$  from  $s_0 = s_{init}$  to  $s_n = goal$  with  $goal \notin \{s_0, \dots, s_{n-1}\}$  and  $\Pi_{\bowtie N} = \bigcup_{n \bowtie N} \Pi_n$ , e.g.,  $\Pi_{\leq N} = \Pi_0 \cup \Pi_1 \cup \dots \cup \Pi_N$  and  $\Pi_{> N} = \Pi_{N+1} \cup \Pi_{N+2} \cup \dots$ .

We first suppose that  $\mathbb{CE}^{\mathfrak{S}} > 0$ . Then,  $z = \mathbb{CE}^{\mathfrak{S}} \cdot \Pr_{s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$ . We pick some  $n_0 \in \mathbb{N}$  such that

$$x_0 = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\bigcirc^{\leq n_0} goal) = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Pi_{\leq n_0}) > 0$$

and such that there is at least one finite  $\mathfrak{S}$ -path  $\pi \in \Pi_{\leq n_0}$  with  $rew(\pi) > 0$ . Lemma E.4 (applied to  $\mathcal{M}$  with the unit-reward function) and Lemma E.5 yield a step bound  $n_\varepsilon \geq n_0$  such that

$$\begin{aligned} \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{T}}(\bigcirc^{\leq n_\varepsilon} (goal \vee fail)) &\geq 1 - \frac{1}{4} \cdot x_0^2 \cdot \varepsilon \cdot \frac{1}{z} \\ \sum_{n=n_\varepsilon+1}^{\infty} \sum_{\pi \in \Pi_n} rew(\pi) \cdot \text{prob}^{\mathfrak{T}}(\pi) &\leq \frac{1}{4} \cdot x_0^2 \cdot \varepsilon \end{aligned}$$

for all schedulers  $\mathfrak{T}$ . Let  $x = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\bigcirc^{\leq n_\varepsilon} goal)$  and

$$\rho = \sum_{n=0}^{n_\varepsilon} \sum_{\pi \in \Pi_n} rew(\pi) \cdot \text{prob}^{\mathfrak{S}}(\pi) = z - \sum_{n=n_\varepsilon+1}^{\infty} \sum_{\pi \in \Pi_n} rew(\pi) \cdot \text{prob}^{\mathfrak{S}}(\pi)$$

Then,  $x_0 \leq x$  and  $\rho \leq z$ . Moreover,  $\rho > 0$  by the choice of  $n_0$  and the requirement  $n_\varepsilon \geq n_0$ . In particular:

$$y^{\mathfrak{T}} \stackrel{\text{def}}{=} \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{T}}(\Pi_{> n_\varepsilon}) < \frac{1}{4} \cdot x^2 \cdot \varepsilon \cdot \frac{1}{\rho}$$

For all non-negative real number  $y, \theta$  with  $y < \frac{1}{4} \cdot x^2 \cdot \varepsilon \cdot \frac{1}{\rho}$  and  $\theta < \frac{1}{4} \cdot x^2 \cdot \varepsilon$  we have:

$$\left| \frac{\rho + \theta}{x + y} - \frac{\rho}{x} \right| = \frac{|\theta \cdot x - \rho \cdot y|}{(x + y) \cdot x} \leq \frac{\theta \cdot x + \rho \cdot y}{x^2} < \frac{\varepsilon}{2}$$

Suppose now that  $\mathfrak{T}$  agrees with  $\mathfrak{S}$  for all paths up to length  $n_\varepsilon$ . Then:

$$\mathbb{CE}^{\mathfrak{S}} = \frac{\rho + \theta^{\mathfrak{S}}}{x + y^{\mathfrak{S}}} \quad \mathbb{CE}^{\mathfrak{T}} = \frac{\rho + \theta^{\mathfrak{T}}}{x + y^{\mathfrak{T}}} \quad \text{where } \theta^{\mathfrak{T}} = \sum_{n > n_\varepsilon} \sum_{\pi \in \Pi_n} rew(\pi) \cdot \text{prob}^{\mathfrak{S}}(\pi)$$

The definition of  $\theta^{\mathfrak{S}}$  is analogous. By the choice of  $n_\varepsilon$  we have  $\theta^{\mathfrak{S}}, \theta^{\mathfrak{T}} < \frac{1}{4} \cdot x^2 \cdot \varepsilon$ . We obtain:

$$|\mathbb{CE}^{\mathfrak{S}} - \mathbb{CE}^{\mathfrak{T}}| \leq \left| \mathbb{CE}^{\mathfrak{S}} - \frac{\rho}{x} \right| + \left| \frac{\rho}{x} - \mathbb{CE}^{\mathfrak{T}} \right| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

It remains to consider the case  $\mathbb{CE}^{\mathfrak{S}} = 0$ . In this case we pick some  $n_\varepsilon \in \mathbb{N}$  such that  $x = \Pr_{s_{init}}^{\mathfrak{S}}(\bigcirc^{\leq n_\varepsilon} goal) > 0$  and

$$\theta^{\mathfrak{T}} = \sum_{n=n_\varepsilon+1}^{\infty} \sum_{\pi \in \Pi_n} \text{rew}(\pi) \cdot \text{prob}^{\mathfrak{T}}(\pi) < x \cdot \varepsilon$$

for all schedulers  $\mathfrak{T}$  (Lemma E.5). If  $\mathfrak{S}$  and  $\mathfrak{T}$  agree on all paths up to length  $n_\varepsilon$  then:

$$\mathbb{CE}^{\mathfrak{T}} = \frac{\theta^{\mathfrak{T}}}{x + y^{\mathfrak{T}}} \leq \frac{\theta^{\mathfrak{T}}}{x} < \varepsilon$$

where  $y^{\mathfrak{T}}$  is as above. ■

### E.3 Optimal reward-based eventually memoryless schedulers

**Proposition E.8 (Turning point).** *There exists  $\mathfrak{R} \in \mathbb{N}$  such that for each scheduler  $\mathfrak{S}$  the following statement holds. If  $\pi$  is a finite  $\mathfrak{S}$ -path from  $s_{init}$  to some state  $s \in S \setminus \{goal, fail\}$  such that  $\text{rew}(\pi) \geq \mathbb{CE}^{\mathfrak{S}} + 1$  and  $\mathbb{CE}_s^{\mathfrak{S} \uparrow \pi} \geq \mathfrak{R}$  then*

$$\mathbb{CE}^{\mathfrak{S}} < \mathbb{CE}^{\mathfrak{S} \triangleleft_\pi \mathfrak{U}}$$

where  $\mathfrak{U}$  is an arbitrary scheduler that maximizes the probability to reach goal from  $s$ . The value  $\mathfrak{R}$  will be called a turning point of  $\mathcal{M}$ .

*Proof.* Recall that  $p_s^{\max} = \Pr_s^{\max}(\Diamond goal)$ . Let

$$p = \min_{\substack{s \in S \\ s \neq fail}} p_s^{\max}$$

The default assumption stating that *goal* is reachable from all states in  $S \setminus \{fail\}$  yields  $p > 0$ . Let  $\lambda \in ]0, 1[$  and  $R_0 \in \mathbb{N}$  be as in Lemma E.4 and  $C$  as in Lemma E.5. Recall that  $C = 1/(1-\lambda)^2$ , which yields  $C > 1 \geq p$ . We pick some  $\mathfrak{R} \geq R_0$  such that  $\mathfrak{R}$  is even and

$$C \cdot \mathfrak{R} \cdot \lambda^{\frac{\mathfrak{R}}{2}} < \frac{p}{2}$$

Let  $\pi$  be a finite  $\mathfrak{S}$ -path from  $s_{init}$  to some state  $s \in S \setminus \{goal, fail\}$  such that  $\text{rew}(\pi)$  is positive and  $\mathbb{CE}_s^{\mathfrak{S} \uparrow \pi} \geq \mathfrak{R}$ . We proceed as follows. We first establish upper bounds for  $\Pr_s^{\mathfrak{T}}(\Diamond goal)$  and  $\mathbb{E}_s^{\mathfrak{T}}(\Diamond goal)$ . These bounds will be used in the second part where we prove  $\mathbb{CE}^{\mathfrak{S}} < \mathbb{CE}^{\mathfrak{S} \triangleleft_\pi \mathfrak{U}}$ .

Let  $\mathfrak{T} = \mathfrak{S} \uparrow \pi$  be the residual scheduler and  $y = \Pr_s^{\mathfrak{T}}(\Diamond goal)$ . The first part of Corollary E.6 yields:

$$\mathbb{E}_s^{\mathfrak{T}} \leq \frac{\mathfrak{R}}{2} \cdot y + C \cdot \frac{\mathfrak{R}}{2} \cdot \lambda^{\frac{\mathfrak{R}}{2}}$$

By assumption we have  $\mathbb{CE}_s^{\mathfrak{T}} \geq \mathfrak{R}$ . By the choice of  $\mathfrak{R}$  we have  $C \cdot \mathfrak{R} \cdot \lambda^{\frac{\mathfrak{R}}{2}} < p/2$ . Hence:

$$y \leq C \cdot \lambda^{\frac{\mathfrak{R}}{2}} < \frac{p}{2} \cdot \lambda^{\frac{\mathfrak{R}}{2}}$$

by the second part of Corollary E.6. The fact that  $C > p$  implies  $C > p/2$ . By the choice of  $\mathfrak{R}$  we obtain:

$$\mathbb{E}_s^{\mathfrak{T}} \leq \frac{p}{2} \cdot \frac{\mathfrak{R}}{2} \cdot \lambda^{\frac{\mathfrak{R}}{2}} + C \cdot \frac{\mathfrak{R}}{2} \cdot \lambda^{\frac{\mathfrak{R}}{2}} \leq C \cdot \mathfrak{R} \cdot \lambda^{\frac{\mathfrak{R}}{2}} < \frac{p}{2}$$

We now compare the conditional expectations of the schedulers  $\mathfrak{S}$  and  $\mathfrak{S} \triangleleft_{\pi} \mathfrak{U}$  where  $\mathfrak{U}$  is a memoryless scheduler maximizing the probabilities to reach the goal state from each state. That is,  $p_t^{\max} = \Pr_t^{\mathfrak{U}}(\Diamond goal) > 0$ .

Let  $r = rew(\pi)$  and  $z = \text{prob}(\pi)$ . By assumption  $r \geq \mathbb{CE}^{\mathfrak{S}} + 1$ . We define:

$$\rho = \mathbb{E}_{s_{init}}^{\mathfrak{S}} - yr \quad \text{and} \quad x = \Pr_{s_{init}}^{\mathfrak{S}}(\Diamond goal) - y$$

Recall that  $s = \text{last}(\pi)$ . Then:

$$\mathbb{CE}^{\mathfrak{S}} = \frac{\rho + z(yr + \mathbb{E}_s^{\mathfrak{T}})}{x + zy}$$

and

$$\mathbb{CE}^{\mathfrak{S} \triangleleft_{\pi} \mathfrak{U}} = \frac{\rho + z(p_s^{\max}r + \mathbb{E}_s^{\mathfrak{U}})}{x + zp_s^{\max}} \geq \frac{\rho + zp_s^{\max}r}{x + zp_s^{\max}}$$

Thus, to prove  $\mathbb{CE}^{\mathfrak{S}} < \mathbb{CE}^{\mathfrak{S} \triangleleft_{\pi} \mathfrak{U}}$ , it suffices to show:

$$\frac{\rho + z(yr + \mathbb{E}_s^{\mathfrak{T}})}{x + zy} < \frac{\rho + zp_s^{\max}r}{x + zp_s^{\max}}$$

We now use the bounds  $\mathbb{E}_s^{\mathfrak{T}} < p/2$  and  $y < p/2 \cdot \lambda^{\frac{\mathfrak{R}}{2}}$  (which yields  $y < p/2$ ) that have been established above and obtain:

$$\begin{aligned} \frac{zp_s^{\max}r - z(yr + \mathbb{E}_s^{\mathfrak{T}})}{zp_s^{\max} - zy} &= \frac{p_s^{\max}r - (yr + \mathbb{E}_s^{\mathfrak{T}})}{p_s^{\max} - y} \\ &= \frac{p_s^{\max}r - yr}{p_s^{\max} - y} - \frac{\mathbb{E}_s^{\mathfrak{T}}}{p_s^{\max} - y} \\ &= r - \frac{\mathbb{E}_s^{\mathfrak{T}}}{p_s^{\max} - y} \\ &\geq r - \frac{\mathbb{E}_s^{\mathfrak{T}}}{p - y} \\ &> r - \frac{\frac{p}{2}}{p - \frac{p}{2}} = r - 1 \end{aligned}$$

By assumption we have  $r = \text{rew}(\pi) \geq \mathbb{CE}^{\mathfrak{S}} + 1$ . Thus,  $r-1 \geq \mathbb{CE}^{\mathfrak{S}}$ . Therefore:

$$\frac{zp_s^{\max}r - z(yr + E_s^{\mathfrak{T}})}{zp_s^{\max} - zr} > r-1 \geq \mathbb{CE}^{\mathfrak{S}}$$

By Lemma E.3 we get:

$$\mathbb{CE}^{\mathfrak{S}} < \frac{\rho + zp_s^{\max}r}{x + zp_s^{\max}} < \frac{zp_s^{\max}r - z(yr + E_s^{\mathfrak{T}})}{zp_s^{\max} - zy}$$

But then  $\mathbb{CE}^{\mathfrak{S}} < (\rho + zp_s^{\max}r)/(x + zp_s^{\max}) \leq \mathbb{CE}^{\mathfrak{S} \triangleleft_{\pi} \mathfrak{U}}$ .  $\blacksquare$

Given a reward-based scheduler  $\mathfrak{S}$  and a state-reward pair  $(s, r) \in S \times \mathbb{N}$  with  $r \geq \mathbb{CE}^{\mathfrak{S}} + 1$  and  $\mathbb{CE}_s^{\mathfrak{S} \uparrow(s, r)} \geq \mathfrak{R}$ , we may applying Proposition E.8 repeatedly to obtain

$$\mathbb{CE}^{\mathfrak{S}} < \mathbb{CE}^{\mathfrak{S} \triangleleft_{(s, r)} \mathfrak{U}}$$

where  $\mathfrak{U}$  is any scheduler that maximizes the probability to reach *goal* from  $s$ . Hence, by Proposition D.1,  $\mathbb{CE}^{\max}$  is the supremum of the values  $\mathbb{CE}^{\mathfrak{S}}$  where  $\mathfrak{S}$  ranges over all deterministic reward-based schedulers with  $\mathbb{CE}^{\mathfrak{S} \uparrow r} < \mathfrak{R}$  for all  $r \geq \mathbb{CE}^{\mathfrak{S}} + 1$ .

**Definition E.9 (Eventually memoryless).** *Let  $\mathfrak{S}$  be a reward-based scheduler.  $\mathfrak{S}$  is called eventually memoryless if there exists  $\wp \in \mathbb{N}$  such that  $\mathfrak{S}(s, r) = \mathfrak{S}(s, \wp)$  for all  $r \geq \wp$ .*

We will show that there exists an optimal reward-based eventually memoryless scheduler  $\mathfrak{S}$  such that for all states  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and each  $r \geq \wp$  we have  $\mathfrak{S} \uparrow r = \mathfrak{M}(s)$  where  $\mathfrak{M}$  is a deterministic memoryless scheduler that maximizes the probability to reach *goal* from all states and the conditional expectations for all those schedulers. We will see that such a scheduler  $\mathfrak{M}$  is computable in polynomial time using linear programming techniques. See Lemma E.14 below.

**Definition E.10 (Additional notations).** As before, let  $p_s^{\max} = \Pr_s^{\max}(\diamond \text{goal})$ . Let  $\text{Act}^{\max}(s)$  denote the set of actions  $\alpha \in \text{Act}(s)$  where

$$p_s^{\max} = \sum_{t \in S} P(s, \alpha, t) \cdot p_t^{\max}$$

Let  $\text{Sched}^{\max}$  denote the class of deterministic schedulers  $\mathfrak{U}$  such that  $\Pr_s^{\mathfrak{U}}(\diamond \text{goal}) = p_s^{\max}$  for all states  $s$ .  $\blacksquare$

It is well-known (see e.g. [33]) that  $\text{Act}^{\max}(s)$  is nonempty and that  $\mathfrak{U}(\pi) \in \text{Act}^{\max}(\text{last}(\pi))$  for each  $\mathfrak{U}$ -path starting in  $s$  and each scheduler  $\mathfrak{U} \in \text{Sched}^{\max}$ . This justifies to regard the sub-MDP  $\mathcal{M}^{\max}$  of  $\mathcal{M}$  that arises by eliminating all state-action pairs  $(s, \beta)$  with  $s \in S$  and  $\beta \notin \text{Act}^{\max}(s)$ . That is, the enabled action of  $s$  as a state of  $\mathcal{M}^{\max}$  are exactly the actions in  $\text{Act}^{\max}(s)$ . Clearly:

$$\Pr_{\mathcal{M}, s}^{\mathfrak{U}}(\diamond \text{goal}) = \Pr_{\mathcal{M}^{\max}, s}^{\mathfrak{U}}(\diamond \text{goal})$$

for each scheduler  $\mathfrak{V}$  for  $\mathcal{M}^{\max}$  and each scheduler  $\mathfrak{U} \in \text{Sched}^{\max}$  is also a scheduler for  $\mathcal{M}^{\max}$ . The reverse direction does not hold in general as  $\mathcal{M}^{\max}$  can have end components that do not contain the goal state, i.e.,  $\Pr_s^{\mathfrak{V}}(\Diamond \text{goal}) < p_s^{\max}$  for some scheduler  $\mathfrak{V}$  for  $\mathcal{M}^{\max}$  is possible. However, such scenarios are impossible because of assumptions (A1) and (A2).

**Lemma E.11.**  *$\text{Sched}^{\max}$  agrees with the set of schedulers for  $\mathcal{M}^{\max}$ . That is, for each scheduler  $\mathfrak{U}$  for  $\mathcal{M}^{\max}$  we have  $\Pr_{\mathcal{M},s}^{\mathfrak{U}}(\Diamond \text{goal}) = p_s^{\max}$  for all states  $s$ .*

*Proof.* Let  $\mathfrak{S}$  be a deterministic memoryless schedulers for  $\mathcal{M}^{\max}$  where  $\Pr_s^{\mathfrak{S}}(\Diamond \text{goal})$  is minimal for all states  $s$ . Let  $q_s = \Pr_s^{\mathfrak{S}}(\Diamond \text{goal})$  and  $\beta_s = \mathfrak{S}(s)$ . Let  $S' = \{s \in S : q_s > 0\}$ . Then, the vector  $(q_s)_{s \in S'}$  is the unique solution of the following linear equation system with variables  $x_s$  for  $s \in S'$ :

$$\begin{aligned} x_s &= \sum_{t \in S} P(s, \beta_s, t) \cdot x_t & \text{for } s \in S' \setminus \{\text{goal}\} \\ x_{\text{goal}} &= 1 \end{aligned}$$

But the vector  $(p_s^{\max})_{s \in S'}$  also solves the above linear equation system. Hence,  $q_s = p_s^{\max}$  for all states  $s \in S'$ .

It remains to show that  $S \setminus \{\text{fail}\} = S'$ . For all states  $s \in S \setminus S'$  we have  $\Pr_s^{\mathfrak{S}}(\Diamond \text{fail}) = 1$  by assumption (A2) and the vector  $(w_s)_{s \in S \setminus S'}$  with  $w_s = 1$  for all  $s \in S \setminus S'$  is the unique solution of the following linear equation system with variables  $y_s$  for  $s \in S \setminus S'$ :

$$\begin{aligned} y_s &= \sum_{t \in S} P(s, \beta_s, t) \cdot y_t & \text{for } s \in S \setminus (S' \cup \{\text{fail}\}) \\ y_{\text{fail}} &= 1 \end{aligned}$$

However, the vector  $(1 - p_s^{\max})_{s \in S \setminus S'}$  also solves the above linear equation system. Hence,  $1 - p_s^{\max} = 1$  and therefore  $p_s^{\max} = 0$  for all  $s \in S \setminus S'$ . But then  $S \setminus S' = \{\text{fail}\}$  by assumption (A1). ■

*Remark E.12.* Obviously, for each scheduler  $\mathfrak{S} \in \text{Sched}^{\max}$  we have:

$$\text{CE}^{\mathfrak{S}} = \frac{E_{s_{\text{init}}}^{\mathfrak{S}}}{p_{s_{\text{init}}}^{\max}}$$

Hence, if  $\mathfrak{S}, \mathfrak{U} \in \text{Sched}^{\max}$  then

$$\text{CE}^{\mathfrak{S}} \geq \text{CE}^{\mathfrak{U}} \quad \text{iff} \quad E_{s_{\text{init}}}^{\mathfrak{S}} \geq E_{s_{\text{init}}}^{\mathfrak{U}}$$

For each  $\mathfrak{S} \in \text{Sched}^{\max}$  and each  $\mathfrak{S}$ -path  $\pi$  from  $s_{\text{init}}$ , the residual schedulers  $\mathfrak{S} \uparrow \pi$  maximize the probabilities to reach *goal* from *last*( $\pi$ ). Hence, we may suppose  $\mathfrak{S} \uparrow \pi \in \text{Sched}^{\max}$ . Thus,  $\text{CE}^{\mathfrak{S}}$  is maximal under all schedulers in  $\text{Sched}^{\max}$  iff

$$E_s^{\mathfrak{S} \uparrow \pi} = \sup \{ E_s^{\mathfrak{U}} : \mathfrak{U} \in \text{Sched}^{\max} \}$$

for all  $\mathfrak{S}$ -paths  $\pi$  from  $s_{\text{init}}$  with  $s = \text{last}(\pi) \neq \text{fail}$ . This follows by the fact that

$$\frac{\rho + \theta}{x + p} \geq \frac{\rho + \zeta}{x + p} \quad \text{iff} \quad \theta \geq \zeta$$

for all real numbers  $\rho, \theta, \zeta, x, p$  with  $x + p > 0$ . In this case, we deal with  $p = \text{prob}(\pi)$ ,  $x = p_{s_{\text{init}}}^{\max} - p$ ,  $\theta = E_s^{\mathfrak{G}\uparrow\pi}$  and  $\rho = E_{s_{\text{init}}}^{\mathfrak{G}} - p\theta$ . Thus,  $\mathbb{CE}^{\mathfrak{G}} = (\rho + \theta)/(x + p)$ . The value  $\zeta$  stands for the possible values  $E_s^{\mathfrak{U}}$  for  $\mathfrak{U} \in \text{Sched}^{\max}$ . ■

**Lemma E.13.** *Let  $\Theta_s = \sup \{ E_s^{\mathfrak{U}} : \mathfrak{U} \in \text{Sched}^{\max} \}$ . Then:*

$$\Theta_s = \max \left\{ \text{rew}(s, \alpha) \cdot p_s^{\max} + \sum_{t \in S} P(s, \alpha, t) \cdot \Theta_t : \alpha \in \text{Act}^{\max}(s) \right\}$$

*Proof.* Clearly, for each state  $s \in S \setminus \{\text{fail}\}$  and  $\alpha \in \text{Act}^{\max}(s)$  and each deterministic scheduler  $\mathfrak{U}$  we have:

$$E_s^{\mathfrak{U}} = \text{rew}(s, \alpha) \cdot p_s^{\max} + \sum_{t \in S} P(s, \alpha, t) \cdot E_t^{\mathfrak{U}\uparrow(s \alpha t)}$$

where  $\alpha = \mathfrak{U}(s)$ . This yields  $\Theta_s \geq \Xi_s$  for all states  $s$  where

$$\Xi_s = \max \left\{ \text{rew}(s, \alpha) \cdot p_s^{\max} + \sum_{t \in S} P(s, \alpha, t) \cdot \Theta_t : \alpha \in \text{Act}^{\max}(s) \right\}$$

It remains to show that  $\Xi_s \leq \Theta_s$  for all states  $s$ . Suppose by contradiction that  $\Theta_s > \Xi_s$  for some state  $s$ . Let  $\varepsilon = (\Theta_s - \Xi_s)/2$ . We pick some deterministic scheduler  $\mathfrak{G} \in \text{Sched}^{\max}$  such that  $E_s^{\mathfrak{G}} > \Theta_s - \varepsilon$ . Hence,  $E_s^{\mathfrak{G}} > \Xi_s$ . For  $\alpha = \mathfrak{G}(s)$ , we get:

$$\Xi_s < E_s^{\mathfrak{G}} \leq \text{rew}(s, \alpha) \cdot p_s^{\max} + \sum_{t \in S} P(s, \alpha, t) \cdot \Theta_t \leq \Xi_s$$

Contradiction. ■

**Lemma E.14.** *There exists a deterministic memoryless scheduler  $\mathfrak{M} \in \text{Sched}^{\max}$  that maximizes the partial expected total reward until reaching goal for all states  $s \in S \setminus \{\text{fail}\}$  under all schedulers  $\mathfrak{U} \in \text{Sched}^{\max}$ , i.e.,*

$$E_s^{\mathfrak{M}} = \max \{ E_s^{\mathfrak{U}} : \mathfrak{U} \in \text{Sched}^{\max} \}$$

*Such a scheduler  $\mathfrak{M}$  and the values  $E_s^{\mathfrak{M}}$  are computable in time polynomial in the size of  $\mathcal{M}^{\max}$ , using the linear program with variables  $\theta_s$  for  $s \in S$  shown in Figure 2.*

By Lemma E.14 and Remark E.12:  $\mathbb{CE}_s^{\mathfrak{M}} = \max \{ \mathbb{CE}_s^{\mathfrak{U}} : \mathfrak{U} \in \text{Sched}^{\max} \}$ .

*Proof.* The linear program in Figure 2 is the same as the one for the maximal (unconditional) total expectation of the MDP  $\mathcal{M}'$  that agrees with  $\mathcal{M}^{\max}$ , but uses the (rational-valued) reward function  $\text{rew}'(s, \alpha) = \text{rew}(s, \alpha) \cdot p_s^{\max}$ . Using standard results for finite MDPs (see e.g. [33]), we get that the linear program has a unique solution. Thus, one proof obligation is to show that  $E_{\mathcal{M}^{\max}, s}^{\max}(\Diamond \text{goal}) = \mathbb{E}_{\mathcal{M}', s}^{\max}$  (“total reward”) for all states  $s$  and to show that optimal schedulers for  $\mathcal{M}'$  (w.r.t. to the total expected reward) are optimal for  $\mathcal{M}^{\max}$  (w.r.t. the partial

Minimize  $\sum_{s \in S} \theta_s$  subject to

- (1)  $\theta_s \geq \text{rew}(s, \alpha) \cdot p_s^{\max} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_t$  for  $s \in S \setminus \{\text{goal}, \text{fail}\}$ ,  $\alpha \in \text{Act}^{\max}(s)$
- (2)  $\theta_{\text{goal}} = \theta_{\text{fail}} = 0$  and  $\theta_s \geq 0$  for  $s \in S \setminus \{\text{goal}, \text{fail}\}$

**Fig. 2.** Linear program for  $\max \{ E_s^{\mathfrak{U}} : \mathfrak{U} \in \text{Sched}^{\max} \}$

expectation until reaching the goal state). We follow here a different approach and present a direct proof that adapts the soundness of the linear program for total expected accumulated rewards in finite MDPs.

Clearly, the vector  $(\Theta_s)_{s \in S}$  defined as Lemma E.13 provides a solution for the constraints (1) and (2) in Figure 2. Moreover, Lemma E.13 implies that for each state  $s \in S \setminus \{\text{goal}, \text{fail}\}$  there is an action  $\beta_s \in \text{Act}^{\max}(s)$  such that

$$\Theta_s = \text{rew}(s, \beta_s) \cdot p_s^{\max} + \sum_{t \in S} P(s, \beta_s, t) \cdot \Theta_t$$

Let  $\mathfrak{M}$  be the deterministic memoryless scheduler for  $\mathcal{M}^{\max}$  given by  $\mathfrak{M}(s) = \beta_s$  for all states  $s \in S \setminus \{\text{goal}, \text{fail}\}$ . By Lemma E.11 we get  $\mathfrak{M} \in \text{Sched}^{\max}$ , i.e.,  $\text{Pr}_s^{\mathfrak{M}}(\diamond \text{goal}) = p_s^{\max}$  for all  $s$ .

The vectors  $(E_s^{\mathfrak{M}})_{s \in S}$  and  $(\Theta_s)_{s \in S}$  solve the following linear equation system with variables  $\zeta_s$  for all states  $s \in S$ :

- (3)  $\zeta_s = \text{rew}(s, \beta_s) \cdot p_s^{\max} + \sum_{t \in S} P(s, \beta_s, t) \cdot \zeta_t$  for  $s \in S \setminus \{\text{goal}, \text{fail}\}$
- (4)  $\zeta_{\text{goal}} = \zeta_{\text{fail}} = 0$

By applying standard arguments for the Markov chain induced by  $\mathfrak{M}$  and using assumption (A2), we obtain that the above linear equation system has a unique solution.<sup>9</sup> This yields:  $\Theta_s = E_s^{\mathfrak{M}}$  for all states  $s \in S$ .

It remains to show that  $\sum_{s \in S} \Theta_s \leq \sum_{s \in S} \rho_s$  for each solution  $(\rho_s)_{s \in S}$  of (1) and (2). We pick a solution  $(\rho_s)_{s \in S}$  of (1) and (2). We first observe that then also the vector with the elements  $\min\{\Theta_s, \rho_s\}$  is a solution of (1) and (2). Hence, we may assume that  $\rho_s \leq \Theta_s$  for all  $s \in S$ .

We now define  $\rho_s^{(0)} = \rho_s$  and for  $n \in \mathbb{N}$ :

$$\rho_s^{(n+1)} = \text{rew}(s, \beta_s) \cdot p_s^{\max} + \sum_{t \in S} P(s, \alpha, t) \cdot \rho_t^{(n)}$$

By induction on  $n$ , we get  $\rho_s^{(0)} \geq \rho_s^{(1)} \geq \rho_s^{(2)} \geq \dots$  for all  $s \in S$  and  $n \geq 0$ . Let

$$\rho_s^* = \lim_{n \rightarrow \infty} \rho_s^{(n)}$$

<sup>9</sup> Note that the linear equation system (3), (4) can be written in the form  $(I - A)\zeta = b$  where  $A$  is the probability matrix of the Markov chain induced by  $\mathfrak{M}$  restricted to the states  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and  $I$  the identity matrix. The vector  $b$  contains the values  $\text{rew}(s, \beta_s) \cdot p_s^{\max}$ ,  $s \in S \setminus \{\text{goal}, \text{fail}\}$ . Assumption (A2) ensures that  $I - A$  is non-singular. This implies the existence of a unique solution of equations (3) and (4).



Clearly, we have  $\rho_s \geq \rho_s^*$  and

$$\rho_s^* = \text{rew}(s, \beta_s) \cdot p_s^{\max} + \sum_{t \in S} P(s, \alpha, t) \cdot \rho_t^*$$

for all states  $s$ . But then the vector  $(\rho_s^*)_{s \in S}$  solves the linear equation system (3) and (4) of above. Again, we can rely on the fact that (3) and (4) have a unique solution, which yields  $\rho_s^* = \Theta_s$  for all states  $s$ . But then  $\rho_s \geq \rho_s^* \geq \Theta_s$  for all  $s$ .

The above shows that the vector  $(\Theta_s)_{s \in S}$  is the unique solution of the linear program shown in Figure 2 and coincides with the vector  $(E_s^{\mathfrak{M}})_{s \in S}$ . ■

**Lemma E.15 (Existence of optimal eventually memoryless schedulers).**  $\mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\max}$  for some deterministic reward-based scheduler  $\mathfrak{S}$  such that  $\text{Pr}_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{S}}(\Diamond \text{goal}) > 0$  and  $\mathfrak{S} \uparrow \wp = \mathfrak{M}$  for some saturation point  $\wp$ .

*Proof.* We define

$$\delta_s = \min \left\{ p_s^{\max} - \sum_{t \in S} P(s, \beta, t) \cdot p_t^{\max} : \beta \in \text{Act}(s) \setminus \text{Act}^{\max}(s) \right\}$$

and with  $S_\delta = \{s \in S : \delta_s > 0\}$ :

$$\delta = \min_{s \in S_\delta} \delta_s$$

If  $S_\delta = \emptyset$  then  $\text{Act}(s) = \text{Act}^{\max}(s)$  for all states  $s$ . In this case, the deterministic memoryless scheduler  $\mathfrak{M}$  as in Lemma E.14 is an optimal scheduler as it maximizes the conditional expectation from every state (see Remark E.12).

In what follows, we suppose that  $S_\delta$  is nonempty, in which case  $\delta$  is positive. Let  $\mathfrak{R}$  be the turning point of  $\mathcal{M}$  as in Proposition E.8. We now define the saturation point  $\wp$  as any natural number satisfying the following constraint:

$$\wp \geq \mathbb{CE}^{\max} + \frac{\mathfrak{R}}{\delta}$$

As  $0 < \delta \leq 1$  we have  $\mathfrak{R}/\delta \geq 1$ . Moreover, we pick a deterministic memoryless scheduler  $\mathfrak{M}$  as in Lemma E.14. Lemma E.16 (see below) shows that for each partial deterministic reward-based scheduler

$$\mathfrak{P} : S \times \{r \in \mathbb{N} : 0 \leq r < \wp\} \rightarrow \text{Act}$$

the scheduler  $\mathfrak{P} \triangleleft_\wp \mathfrak{M}$  is optimal among all schedulers  $\mathfrak{P} \triangleleft_\wp \mathfrak{T}$  where  $\mathfrak{T}$  ranges over all schedulers and where optimality is understood with respect to conditional expectations. Note that the scheduler  $\mathfrak{P} \triangleleft_\wp \mathfrak{M}$  is reward-based eventually memoryless with saturation point  $\wp$ . Since there are only finitely many partial schedulers  $\mathfrak{P}$ , this completes the proof of Lemma E.15. ■

**Lemma E.16.**  $\mathbb{CE}^{\mathfrak{P} \triangleleft_\wp \mathfrak{M}} \geq \mathbb{CE}^{\mathfrak{P} \triangleleft_\wp \mathfrak{T}}$  for each scheduler  $\mathfrak{T}$  and each partial scheduler  $\mathfrak{P} : S \times \{r \in \mathbb{N} : 0 \leq r < \wp\} \rightarrow \text{Act}$ .

*Proof.* We first prove the following claim.

*Claim.* Suppose we are given two schedulers  $\mathfrak{S}$  and  $\mathfrak{T}$  that agree for all but the extensions of some finite path  $\pi$  starting in  $s_{init}$  where  $\pi$  is both a  $\mathfrak{S}$ -path and a  $\mathfrak{T}$ -path from  $s_{init}$ , i.e.,  $\mathfrak{S}\uparrow\varrho = \mathfrak{T}\uparrow\varrho$  for each finite path  $\varrho$  where  $\pi$  is not a prefix of  $\pi$ . Let  $s = last(\pi)$  and suppose  $rew(\pi) \geq \wp$ . Then:

$$\Pr_s^{\mathfrak{S}\uparrow\pi}(\Diamond goal) = p_s^{\max} > \Pr_s^{\mathfrak{T}\uparrow\pi}(\Diamond goal) \text{ implies } \mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{S}} \geq \mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{T}}$$

*Proof of the claim.* We provide the proof of the claim for deterministic schedulers. The argument for randomized schedulers is similar and omitted here as randomized schedulers are irrelevant for our purposes by Proposition D.1. Let  $s = last(\pi)$ ,  $r = rew(\pi)$ ,  $w = \text{prob}(\pi)$  and  $\mathfrak{V} = \mathfrak{T}\uparrow\pi$  and  $\mathfrak{U} = \mathfrak{S}\uparrow\pi$ . The assumption  $\Pr_s^{\mathfrak{U}}(\Diamond goal) = p_s^{\max}$  yields that  $\Pr_t^{\mathfrak{U}\uparrow\varrho}(\Diamond goal) = p_t^{\max}$  for each finite  $\mathfrak{U}$ -path with  $s = first(\varrho)$  and  $t = last(\varrho)$ . We may assume w.l.o.g. that  $\pi$  is minimal with respect to the above property, i.e.,  $\mathfrak{U}(s) \neq \mathfrak{V}(s)$ . Then,  $\mathbb{CE}^{\mathfrak{S}}$  and  $\mathbb{CE}^{\mathfrak{T}}$  have the following form:

$$\mathbb{CE}^{\mathfrak{S}} = \frac{\rho + wpr + wE_s^{\mathfrak{U}}}{x + wp} \quad \text{and} \quad \mathbb{CE}^{\mathfrak{T}} = \frac{\rho + wyr + wE_s^{\mathfrak{V}}}{x + wy}$$

where  $p = p_s^{\max}$  and  $y = \Pr_s^{\mathfrak{V}}(\Diamond goal)$ . Then,  $y < p$  by assumption. The values  $\rho$  and  $x$  are given by

$$\rho = \sum_{\varrho} rew(\varrho) \cdot \text{prob}(\varrho) \quad \text{and} \quad x = \sum_{\varrho} \text{prob}(\varrho)$$

where  $\varrho$  ranges over all  $\mathfrak{S}$ -paths from  $s_{init}$  to  $goal$  where  $\pi$  is not a prefix of  $\varrho$ . These paths are also  $\mathfrak{T}$ -paths. We now rely on Lemma E.3, which yields

$$\mathbb{CE}^{\mathfrak{S}} \geq \mathbb{CE}^{\mathfrak{T}} \quad \text{iff} \quad \frac{(pr + E_s^{\mathfrak{U}}) - (yr + E_s^{\mathfrak{V}})}{p - y} \geq \mathbb{CE}^{\mathfrak{S}}$$

Thus, the task is to show that

$$r(p - y) + E_s^{\mathfrak{U}} - E_s^{\mathfrak{V}} \geq \mathbb{CE}^{\mathfrak{S}}(p - y)$$

As  $r \geq \wp$  and  $\mathbb{CE}^{\mathfrak{S}} \leq \mathbb{CE}^{\max}$  and by the choice of  $\wp$  we have:

$$r - \mathbb{CE}^{\mathfrak{S}} \geq \wp - \mathbb{CE}^{\max} \geq \frac{\mathfrak{R}}{\delta}$$

Recall that  $\mathfrak{R}$  denotes the turning point of  $\mathcal{M}$ . We may assume w.l.o.g. that

$$\frac{E_s^{\mathfrak{V}}}{y} = \mathbb{CE}_s^{\mathfrak{V}} = \mathbb{CE}^{\mathfrak{T}\uparrow\pi} < \mathfrak{R}$$

as otherwise the claim follows immediately by Proposition E.8. But this yields:

$$E_s^{\mathfrak{V}} < \mathfrak{R}y < \mathfrak{R}p$$

As  $\mathfrak{U}(s) \neq \mathfrak{V}(s)$  and  $y < p = p_s^{\max}$  we have  $p - y \geq \delta$ . (Here, we use the assumption that  $\mathfrak{U}$  and  $\mathfrak{V}$  are deterministic.) But then  $(p - y)/\delta \geq 1$  and therefore:

$$\begin{aligned} E_s^{\mathfrak{V}} - E_s^{\mathfrak{U}} &\leq E_s^{\mathfrak{V}} \leq \mathfrak{R}p \leq \mathfrak{R} \leq \frac{p - y}{\delta} \cdot \mathfrak{R} \\ &= \frac{\mathfrak{R}}{\delta}(p - y) \leq (r - \mathbb{CE}^{\mathfrak{S}})(p - y) \end{aligned}$$

Hence,  $r(p - y) + E_s^{\mathfrak{U}} - E_s^{\mathfrak{V}} \geq \mathbb{CE}^{\mathfrak{S}}(p - y)$ . This completes the proof of the claim.

*Proof of Lemma E.16.* Let  $S'$  denote the set of states  $s \in S$  such that  $s$  is the last state of some finite  $\mathfrak{P}$ -path  $\pi$  such that  $r = \text{rew}(\pi) < \wp$  and  $\text{rew}(s, \mathfrak{P}(s, r)) \geq \wp$ . Lemma E.14 yields:

$$\mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{M}} \geq \mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{T}}$$

for each scheduler  $\mathfrak{T}$  where  $\Pr_s^{\mathfrak{T}}(\Diamond \text{goal}) = p_s^{\max}$  for all  $s \in S'$ . In the sequel, we address the case where  $\mathfrak{T}$  is a scheduler with  $\Pr_s^{\mathfrak{T}}(\Diamond \text{goal}) < p_s^{\max}$  for some state  $s \in S'$ . Let  $\pi_1, \pi_2, \pi_3, \dots$  be an enumeration of all  $\mathfrak{T}$ -paths such that (i)  $\text{rew}(\pi_i) \geq \wp$  and (ii)  $\Pr_{s_i}^{\mathfrak{T} \uparrow \pi}(\Diamond \text{goal}) < p_{s_i}^{\max}$  where  $s_i = \text{last}(\pi_i)$  and such that no proper prefix of  $\pi_i$  enjoys these two properties (i) and (ii). Furthermore, we suppose that  $|\pi_1| \leq |\pi_2| \leq \dots$ . We successively apply the claim to obtain a sequence of schedulers  $\mathfrak{T}_0 = \mathfrak{T}, \mathfrak{T}_1, \mathfrak{T}_2, \dots$  such that  $\mathfrak{T}_{i+1} = \mathfrak{T}_i \triangleleft_{\pi_i} \mathfrak{M}$  and

$$\mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{T}_0} \leq \mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{T}_1} \leq \mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{T}_2} \leq \dots$$

Moreover, the limit of the schedulers  $\mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{T}_i}$  is  $\mathfrak{P} \triangleleft_{\wp} \mathfrak{M}$ . Proposition E.7 then yields  $\mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{P} \triangleleft_{\wp} \mathfrak{M}}$ . ■

Obviously, Lemma E.16 implies  $\mathbb{CE}^{\mathfrak{S} \triangleleft_{\wp} \mathfrak{M}} \geq \mathbb{CE}^{\mathfrak{S}}$  for each scheduler  $\mathfrak{S}$  as stated in part (a) of Proposition E.1.

## F Computing a saturation point

Although the proof presented in Appendix E is constructive, the constructed saturation point can be very large. We now present a simple method for generating a smaller saturation point. The rough idea is make use of the observation made in Appendix E stating that optimal schedulers eventually behave as the scheduler  $\mathfrak{M}$ . Here and in what follows,  $\mathfrak{M}$  is a deterministic memoryless scheduler for  $\mathcal{M}$  that maximizes the probability to reach *goal* from each state and whose conditional expectation is maximal under all those schedulers (see Appendix E). The idea is now to compute a saturation point  $\wp$  as the smallest reward value from which on  $\mathfrak{M}$  is better than other schedulers.

Let  $\theta_s = E_{\mathcal{M},s}^{\mathfrak{M}}$  and  $y_s = \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond \text{goal})$ . Thus,  $y_s = p_s^{\max} = \Pr_{\mathcal{M},s}^{\max}(\Diamond \text{goal})$ . For each state-action pair  $(s, \alpha)$  with  $\alpha \in \text{Act}(s)$  we define:

$$y_{s,\alpha} = \sum_{t \in S} P(s, \alpha, t) \cdot y_t \quad \text{and} \quad \theta_{s,\alpha} = \text{rew}(s, \alpha) \cdot y_{s,\alpha} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_t$$

By the choice of  $\mathfrak{M}$ ,  $y_{s,\alpha} \leq y_s$ , and  $\theta_{s,\alpha} \leq \theta_s$  if  $y_{s,\alpha} = y_s$ .

In what follows, we suppose that there is at least one state-action pair  $(s, \alpha)$  with  $y_{s,\alpha} < y_s$  as otherwise the scheduler  $\mathfrak{M}$  maximizes the conditional expectation in  $\mathcal{M}$  (see Lemma E.14). We now define:

$$\wp = \max \{ \lceil \mathbb{CE}^{\text{ub}} - D \rceil, 0 \}$$

where

$$D = \min \left\{ \frac{\theta_s - \theta_{s,\alpha}}{y_s - y_{s,\alpha}} : s \in S, \alpha \in \text{Act}(s), y_{s,\alpha} < y_s \right\}$$

and  $\mathbb{CE}^{\text{ub}}$  is an upper bound for  $\mathbb{CE}^{\text{max}}$  (e.g., the one computed by the algorithm presented in Section C.4).

**Proposition F.1.** *The computed value  $\wp$  is a saturation point for  $\mathcal{M}$ , i.e.,  $\mathbb{CE}^{\mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}$  for each scheduler  $\mathfrak{T}$  with  $\text{Pr}_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T}}(\Diamond \text{goal}) > 0$ .*

The remainder of this section is concerned with the proof of Prop. F.1. Let  $\wp_0$  be some other saturation point, e.g., the one obtained using Lemma E.15 and Lemma E.16. If  $\wp \geq \wp_0$ , then  $\wp$  is obviously a saturation point as well. In what follows, we suppose  $\wp < \wp_0$ . As  $(\mathfrak{T} \triangleleft_{\wp_0}) \triangleleft_{\wp} \mathfrak{M} = \mathfrak{T} \triangleleft_{\wp} \mathfrak{M}$ , it suffices to consider schedulers  $\mathfrak{T}$  that behave as  $\mathfrak{M}$  for all paths  $\pi$  with  $\text{rew}(\pi) \geq \wp_0$ . Furthermore, it suffices to consider reward-based schedulers. In the sequel, let  $\text{Sched}'$  denote the set of reward-based schedulers  $\mathfrak{T}$  for  $\mathfrak{M}$  such that  $\mathfrak{T}(\pi) = \mathfrak{M}(\text{last}(\pi))$  for all paths  $\pi$  with  $\text{rew}(\pi) \geq \wp_0$ . So, the task is to show that  $\mathbb{CE}^{\mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}$  for each scheduler  $\mathfrak{T} \in \text{Sched}'$  with  $\text{Pr}_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T}}(\Diamond \text{goal}) > 0$ .

**Lemma F.2.**  $\theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s \geq \theta_{s,\alpha} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_{s,\alpha}$  for all states  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and all actions  $\alpha \in \text{Act}(s)$ .

*Proof.* By the definition of  $\wp$ , for all  $s \in S \setminus \{\text{goal}, \text{fail}\}$ ,  $\alpha \in \text{Act}(s)$  with  $y_{s,\alpha} < y_s$  we have:

$$\wp \geq \mathbb{CE}^{\text{ub}} - D \geq \mathbb{CE}^{\text{ub}} - \frac{\theta_s - \theta_{s,\alpha}}{y_s - y_{s,\alpha}}$$

and therefore:

$$\theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s \geq \theta_{s,\alpha} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_{s,\alpha}$$

If  $y_s = y_{s,\alpha}$  then  $\theta_s \geq \theta_{s,\alpha}$  (by the choice of  $\mathfrak{M}$ ). The case  $y_s < y_{s,\alpha}$  is impossible as  $\mathfrak{M}$  maximizes the probabilities to reach *goal*. This yields the claim. ■

**Lemma F.3.**  $\theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s \geq E_{\mathcal{M}, s}^{\mathfrak{T}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \text{Pr}_{\mathcal{M}, s}^{\mathfrak{T}}(\Diamond \text{goal})$  for all schedulers  $\mathfrak{T} \in \text{Sched}'$ .

*Proof.* The idea is to define a new MDP  $\mathcal{N}$  that simulates  $\mathcal{M}$  in such a way that the value  $E_{\mathcal{M}, s}^{\mathfrak{T}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \text{Pr}_{\mathcal{M}, s}^{\mathfrak{T}}(\Diamond \text{goal})$  equals the expected accumulated reward until reaching *final* from  $s$  in  $\mathcal{N}$  under scheduler  $\mathfrak{T} \in \text{Sched}'$ . The new MDP  $\mathcal{N}$  operates in two modes and extends  $\mathcal{M}$  by a new trap state *final*. It tracks

the accumulated reward until the moment the accumulated reward surpasses  $\wp_0$ . From that point on,  $\mathcal{N}$  behaves according to  $\mathfrak{M}$ . The accumulated reward of a path in  $\mathcal{M}$  that has surpassed  $\wp_0$  in the last step is bound by:

$$N = \wp_0 + \max_{s, \alpha} \text{rew}(s, \alpha)$$

That is, if  $\pi$  is a finite path in  $\mathcal{M}$  with  $\text{rew}(\pi) \geq \wp_0$  and  $\text{rew}(\pi') < \wp_0$  for all proper prefixes of  $\pi$  then  $\text{rew}(\pi) < N$ .

Formally,  $\mathcal{N}$  is a MDP with negative and positive reward. Its state space is:

$$S_{\mathcal{N}} = S_1 \cup S_2 \cup \{final\}$$

where  $S_1 = S \times \{0, \dots, \wp_0 - 1\}$  (first mode) and  $S_2 = S \times S \times \{\wp_0, \dots, N\}$  (second mode). Intuitively, the pairs  $\langle s, r \rangle \in S_1$  in the first mode represent the current state  $s$  and the accumulated reward  $r$ , while the triples  $\langle s, t, r \rangle$  used for the states in the second mode represent the current state  $s$ , the state  $t$  where the switch to the second mode occurred and the accumulated reward  $r$  until the switch. The auxiliary state *final* is a trap. The initial state of  $\mathcal{N}$  is  $\langle s_{init}, 0 \rangle$ .

The action set is  $Act_{\mathcal{N}} = Act \cup \{\tau\}$ . The transition probabilities for the states in the first mode are as follows. Let  $s \in S \setminus \{goal, fail\}$ ,  $r \in \{0, 1, \dots, \wp_0 - 1\}$ ,  $\alpha \in Act(s)$  and  $r' = r + \text{rew}(s, \alpha)$ . Then:

$$\begin{aligned} P_{\mathcal{N}}(\langle s, r \rangle, \alpha, \langle t, r' \rangle) &= P(s, \alpha, t) \quad \text{if } r' < \wp_0 \\ P_{\mathcal{N}}(\langle s, r \rangle, \alpha, \langle t, t, r' \rangle) &= P(s, \alpha, t) \quad \text{if } r' = r \geq \wp_0 \end{aligned}$$

and  $\text{rew}_{\mathcal{N}}(\langle s, r \rangle, \alpha) = \text{rew}(s, \alpha)$ . In the second mode  $\mathcal{N}$  behaves according to  $\mathfrak{M}$ . That is, if  $s \in S \setminus \{goal, fail\}$ ,  $r \in \{\wp_0, \dots, N\}$  and  $\alpha = \mathfrak{M}(s)$  then:

$$P_{\mathcal{N}}(\langle s, s', r \rangle, \alpha, \langle t, s', r \rangle) = P(s, \alpha, t)$$

and  $\text{rew}_{\mathcal{N}}(\langle s, s', r \rangle, \alpha) = \text{rew}(s, \alpha)$ . Thus,  $Act_{\mathcal{N}}(\langle s, r \rangle) = Act(s)$  for the states in the first mode, while  $Act_{\mathcal{N}}(\langle s, s', r \rangle) = \{\mathfrak{M}(s)\}$  for the states in the second mode.

The goal and fail states in both modes have  $\tau$ -transitions to the final state, and no other actions is enabled in the goal and fail states. We first consider the goal states:

$$P_{\mathcal{N}}(\langle goal, r \rangle, \tau, final) = P_{\mathcal{N}}(\langle goal, s, r \rangle, \tau, final) = 1$$

and

$$\text{rew}_{\mathcal{N}}(\langle goal, r \rangle, \tau) = \text{rew}_{\mathcal{N}}(\langle goal, s, r \rangle, \tau) = -(\mathbb{CE}^{\text{ub}} - \wp)$$

Thus, if  $\pi$  is a path from  $s$  to *goal* in  $\mathcal{M}$  then for the lifted path  $\pi_{\mathcal{N}}$  from  $\langle s, 0 \rangle$  in  $\mathcal{N}$  we have  $\text{rew}_{\mathcal{N}}(\pi_{\mathcal{N}}) = \text{rew}(\pi) - (\mathbb{CE}^{\text{ub}} - \wp)$ . For the fail states, we want to make sure that the partial expectation for the accumulated reward from the initial state to *final* via a fail state is 0 under all schedulers. For this purpose, we define:

$$P_{\mathcal{N}}(\langle fail, r \rangle, \tau, final) = 1, \quad \text{rew}_{\mathcal{N}}(\langle fail, r \rangle, \tau) = -r$$

This ensures that all paths  $\pi$  in  $\mathcal{N}$  from some state  $\langle s, 0 \rangle$  to *final* via a fail state in the first mode have reward 0. For the fail states in the second mode, we define:

$$P_{\mathcal{N}}(\langle fail, s, r \rangle, \tau, final) = 1, \quad \text{rew}_{\mathcal{N}}(\langle fail, s, r \rangle, \tau) = -r - \frac{E_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail)}{\Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail)}$$

provided that  $\Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail) > 0$ . If  $\Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail) = 1$  then state  $\langle fail, s, r \rangle$  is not reachable and the transition probabilities and reward for  $\langle fail, s, r \rangle$  are irrelevant.

Obviously, there is a one-to-one correspondence between the schedulers  $\mathfrak{T}$  for  $\mathcal{M}$  that belong to  $Sched'$  and the schedulers for  $\mathcal{N}$ . For all schedulers  $\mathfrak{T} \in Sched'$  we have  $\Pr_{\mathcal{N},s}^{\mathfrak{T}}(\Diamond final) = 1$ . For all states  $s \in S$  and all schedulers  $\mathfrak{T} \in Sched'$ :

$$E_{\mathcal{N},\langle s,s',r \rangle}^{\mathfrak{T}}(\Diamond Fail) = E_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail) = \sum_{\pi \in \Pi_s} \text{rew}(\pi) \cdot \text{prob}(\pi)$$

where  $\Pi_s$  denotes the set of finite  $\mathfrak{M}$ -paths  $\pi$  in  $\mathcal{M}$  with  $first(\pi) = s$  and  $last(\pi) = fail$ . Clearly,  $\sum_{\pi \in \Pi_s} \text{prob}(\pi) = \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail)$ . The partial expectation of all paths from  $\langle s, s', r \rangle$  to *final* via some fail state under each scheduler  $\mathfrak{T} \in Sched'$  is:

$$\begin{aligned} & E_{\mathcal{N},\langle s,s',r \rangle}^{\mathfrak{T}}(\Diamond \text{"final via Fail"}) \\ &= \sum_{\pi \in \Pi_s} \left( \text{rew}(\pi) - r - \frac{E_{\mathcal{M},s'}^{\mathfrak{M}}(\Diamond fail)}{\Pr_{\mathcal{M},s'}^{\mathfrak{M}}(\Diamond fail)} \right) \cdot \text{prob}(\pi) \\ &= E_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail) - \left( r + \frac{E_{\mathcal{M},s'}^{\mathfrak{M}}(\Diamond fail)}{\Pr_{\mathcal{M},s'}^{\mathfrak{M}}(\Diamond fail)} \right) \cdot \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail) \end{aligned}$$

where *Fail* denotes the set of all fail states (in either mode) and where we suppose  $\Pr_{\mathcal{M},s'}^{\mathfrak{M}}(\Diamond fail) > 0$ . With  $s = s'$  we get:

$$E_{\mathcal{N},\langle s,s,r \rangle}^{\mathfrak{T}}(\Diamond \text{"final via Fail"}) = -r \cdot \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail)$$

Therefore:

$$\begin{aligned} & E_{\mathcal{N},\langle s,s,r \rangle}^{\mathfrak{T}}(\Diamond final) \\ &= E_{\mathcal{M},s}^{\mathfrak{M}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond goal) - r \cdot \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond fail) \\ &= \theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s - r \cdot (1 - y_s) \end{aligned}$$

For the states in the first mode:

$$E_{\mathcal{N},\langle s,r \rangle}^{\mathfrak{T}}(\Diamond final) = E_{\mathcal{M},s}^{\mathfrak{T}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \Pr_{\mathcal{M},s}^{\mathfrak{T}}(\Diamond goal) - r \cdot \Pr_{\mathcal{M},s}^{\mathfrak{T}}(\Diamond fail)$$

We get for the special case  $r = 0$ :

$$E_{\mathcal{N},\langle s,0 \rangle}^{\mathfrak{T}}(\Diamond final) = E_{\mathcal{M},s}^{\mathfrak{T}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \Pr_{\mathcal{M},s}^{\mathfrak{T}}(\Diamond goal)$$

for all states  $s \in S$  and all schedulers  $\mathfrak{T} \in \text{Sched}'$ . Moreover, for  $\mathfrak{T} = \mathfrak{M}$ , we get for the states in the first mode:

$$\begin{aligned} \mathbb{E}_{\mathcal{N}, \langle s, r \rangle}^{\mathfrak{M}}(\Diamond final) &= \mathbb{E}_{\mathcal{M}, s}^{\mathfrak{M}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \Pr_{\mathcal{M}, s}^{\mathfrak{M}}(\Diamond goal) - r \cdot \Pr_{\mathcal{M}, s}^{\mathfrak{M}}(\Diamond fail) \\ &= \theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s - r \cdot (1 - y_s) \end{aligned}$$

for all states  $s \in S$ . Note that  $\Pr_{\mathcal{M}, s}^{\max}(\Diamond final) = 1 - \Pr_{\mathcal{M}, s}^{\mathfrak{M}}(\Diamond goal) = 1 - y_s$ .

Let  $f : \mathbb{R}^{S_{\mathcal{N}}} \rightarrow \mathbb{R}^{S_{\mathcal{N}}}$  denote the fixed point operator for the maximal (unconditional) expected accumulated reward until *final* in  $\mathcal{N}$ . If  $\phi = (\phi_{\tilde{s}})_{\tilde{s} \in S_{\mathcal{N}}}$  then

$$f(\phi) = (f_{\tilde{s}}(\phi))_{\tilde{s} \in S_{\mathcal{N}}}$$

where  $f_{final}(\phi) = 0$  and for  $\tilde{s} \in S_{\mathcal{N}} \setminus \{final\}$ :

$$f_{\tilde{s}}(\phi) = \max \{ f_{\tilde{s}, \alpha}(\phi) : \alpha \in \text{Act}_{\mathcal{N}}(\tilde{s}) \}$$

where the function  $f_{\tilde{s}, \alpha} : \mathbb{R}^{S_{\mathcal{N}}} \rightarrow \mathbb{R}$  is given by:

$$f_{\tilde{s}, \alpha}(\phi) = \text{rew}_{\mathcal{N}}(\tilde{s}, \alpha) + \sum_{\tilde{t} \in S_{\mathcal{N}}} P_{\mathcal{N}}(\tilde{s}, \alpha, \tilde{t}) \cdot \phi_{\tilde{t}}$$

We now consider the vector  $\phi^* = (\phi_{\tilde{s}}^*)_{\tilde{s} \in S_{\mathcal{N}}}$  where

$$\phi_{\tilde{s}}^* = \mathbb{E}_{\mathcal{N}, \tilde{s}}^{\mathfrak{M}}(\Diamond final)$$

As  $\mathcal{N}$  in the second mode has no nondeterministic choices and behaves as  $\mathfrak{M}$ , we have  $f_{\tilde{s}}(\phi^*) = \phi_{\tilde{s}}^*$  for all states  $\tilde{s}$  in the second mode of  $\mathcal{N}$ . For the states  $\langle s, r \rangle$  of the first mode, we have (see above):

$$\phi_{\langle s, r \rangle}^* = \theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s - r \cdot (1 - y_s)$$

We now show that  $\phi_{\langle s, r \rangle}^* \geq f_{\langle s, r \rangle, \alpha}(\phi^*)$  for each action  $\alpha \in \text{Act}(s) = \text{Act}_{\mathcal{N}}(\langle s, r \rangle)$ . Let  $k = \text{rew}(s, \alpha) = \text{rew}_{\mathcal{N}}(\langle s, r \rangle, \alpha)$ . In the following calculation, we suppose  $r+k < \wp_0$ . The calculation for mode switches (i.e.,  $r+k \geq \wp_0$ ) is similar and omitted here.

$$\begin{aligned} f_{\langle s, r \rangle, \alpha}(\phi^*) &= k + \sum_{t \in S} P_{\mathcal{N}}(\langle s, r \rangle, \alpha, \langle t, r+k \rangle) \cdot \phi_{\langle t, r+k \rangle}^* \\ &= k + \sum_{t \in S} P(s, \alpha, t) \cdot (\theta_t - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_t - (r+k) \cdot (1 - y_t)) \\ &= ky_s + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_t - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \sum_{t \in S} P(s, \alpha, t) \cdot y_t \\ &\quad + k(1 - y_s) - (r+k) \cdot \sum_{t \in S} P(s, \alpha, t) \cdot (1 - y_t) \\ &= \theta_{s, \alpha} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_{s, \alpha} + k(1 - y_{s, \alpha}) - (r+k)(1 - y_{s, \alpha}) \\ &= \theta_{s, \alpha} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_{s, \alpha} - r(1 - y_{s, \alpha}) \end{aligned}$$

If  $\alpha = \mathfrak{M}(s)$  then  $\theta_s = \theta_{s,\alpha}$  and  $y_s = y_{s,\alpha}$ . Hence,  $f_{\langle s,r \rangle, \mathfrak{M}(s)}(\phi^*) = \phi_{\langle s,r \rangle}^*$ . If  $\alpha \in \text{Act}(s) \setminus \{\mathfrak{M}(s)\}$  then  $y_s \geq y_{s,\alpha}$  and

$$\theta_{s,\alpha} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_{s,\alpha} \leq \theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s$$

by Lemma F.2. Hence:

$$f_{\langle s,r \rangle}(\phi^*) = f_{\langle s,r \rangle, \mathfrak{M}(s)}(\phi^*) = \phi_{\langle s,r \rangle}^*$$

This yields  $f(\phi^*) = \phi^*$ . By the results of [14], the vector

$$(\mathbb{E}_{\mathcal{N}, \tilde{s}}^{\max}(\Diamond \text{final}))_{\tilde{s} \in S_{\mathcal{N}}}$$

is the unique fixpoint of  $f$ .<sup>10</sup> Hence,  $\phi_{\tilde{s}}^* = \mathbb{E}_{\mathcal{N}, \tilde{s}}^{\max}(\Diamond \text{final})$  for all states  $\tilde{s} \in S_{\mathcal{N}}$ . That is, scheduler  $\mathfrak{M}$  maximizes the (unconditional) accumulated reward until reaching the final state in  $\mathcal{N}$ . That is,  $\mathbb{E}_{\mathcal{N}, \tilde{s}}^{\mathfrak{M}}(\Diamond \text{final}) \geq \mathbb{E}_{\mathcal{N}, \tilde{s}}^{\mathfrak{T}}(\Diamond \text{final})$  for all schedulers  $\mathfrak{T}$  for  $\mathcal{N}$  and all states  $\tilde{s}$  in  $\mathcal{N}$ . But then:

$$\begin{aligned} \theta_s - (\mathbb{CE}^{\text{ub}} - \wp) \cdot y_s &= \mathbb{E}_{\mathcal{N}, \langle s, 0 \rangle}^{\mathfrak{M}}(\Diamond \text{final}) \\ &\geq \mathbb{E}_{\mathcal{N}, \langle s, 0 \rangle}^{\mathfrak{T}}(\Diamond \text{final}) \\ &= \mathbb{E}_{\mathcal{M}, s}^{\mathfrak{T}} - (\mathbb{CE}^{\text{ub}} - \wp) \cdot \Pr_{\mathcal{M}, s}^{\mathfrak{T}}(\Diamond \text{goal}) \end{aligned}$$

for all states  $s$  in  $\mathcal{M}$  and all schedulers  $\mathfrak{T} \in \text{Sched}'$ .  $\blacksquare$

**Lemma F.4.** *If  $\mathfrak{T} \in \text{Sched}'$  with  $\Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T}}(\Diamond \text{goal}) > 0$  then  $\mathbb{CE}^{\mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}$ .*

*Proof.* Let  $\Gamma$  denote the set of  $\mathfrak{T}$ -paths  $\pi$  in  $\mathcal{M}$  from  $s_{\text{init}}$  to  $\text{goal}$  with  $\text{rew}(\pi) < \wp$ . Let

$$x = \sum_{\pi \in \Gamma} \text{prob}(\pi), \quad \rho = \sum_{\pi \in \Gamma} \text{rew}(\pi) \cdot \text{prob}(\pi)$$

Then,  $x = \Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T}}(\Diamond^{< \wp} \text{goal}) = \Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}(\Diamond^{< \wp} \text{goal})$  and

$$\rho = \sum_{r=0}^{\wp-1} \Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T}}(\Diamond^{=r} \text{goal}) \cdot r = \sum_{r=0}^{\wp-1} \Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}(\Diamond^{=r} \text{goal}) \cdot r$$

Thus, the conditional expectations  $\mathbb{CE}^{\mathfrak{T}}$  and  $\mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}$  have the form

$$\mathbb{CE}^{\mathfrak{T}} = \frac{\rho + \zeta}{x + z} \quad \text{and} \quad \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}} = \frac{\rho + \theta}{x + y}$$

where  $z = \Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T}}(\Diamond^{\geq \wp} \text{goal})$  and  $y = \Pr_{\mathcal{M}, s_{\text{init}}}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}(\Diamond^{\geq \wp} \text{goal})$  and  $\zeta$  and  $\theta$  are the corresponding partial expectations.

<sup>10</sup> [14] considers the fixed point operator for minimal (unconditional) expected accumulated rewards in MDPs. However, [14] treats MDPs that might have negative and positive reward values. Thus, by multiplying all rewards with  $-1$  the results of [14] carry over to maximal expectations.



For  $r \in \mathbb{N}$ ,  $r \geq \wp$ , let  $\Pi_{s,r}$  denote the set of  $\mathfrak{T}$ -paths  $\pi$  from  $s_{init}$  to  $s$  with  $\wp \leq \text{rew}(\pi)$  and such that  $\wp > \text{rew}(\pi')$  for all proper prefixes  $\pi'$  of  $\pi$ . Thus,  $\Pi_{s,r} = \emptyset$  if  $r > N \stackrel{\text{def}}{=} \wp + \max_{s,\alpha} \text{rew}(s, \alpha)$ . Let  $p_{s,r} = \sum_{\pi \in \Pi_{s,r}} \text{prob}(\pi)$ . Then:

$$y = \sum_{s \in S} \sum_{r=\wp}^N p_{s,r} \cdot \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\diamond \text{goal}) = \sum_{s \in S} \sum_{r=\wp}^N p_{s,r} \cdot y_s$$

Similarly:

$$z = \sum_{s \in S} \sum_{r=\wp}^N p_{s,r} \cdot z_{s,r} \quad \text{where} \quad z_{s,r} = \Pr_{\mathcal{M},s}^{\mathfrak{T} \uparrow r}(\diamond \text{goal})$$

and

$$\theta = \sum_{s \in S} \sum_{r=\wp}^N p_{s,r} \cdot (\mathbb{E}_{\mathcal{M},s}^{\mathfrak{M}} + r \cdot y_s) \quad \zeta = \sum_{s \in S} \sum_{r=\wp}^N p_{s,r} \cdot (\mathbb{E}_{\mathcal{M},s}^{\mathfrak{T} \uparrow r} + r \cdot z_{s,r})$$

Note that  $y_s \geq z_{s,r}$  as  $y_s = \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\diamond \text{goal}) = \Pr_{\mathcal{M},s}^{\max}(\diamond \text{goal})$ . Using Lemma F.3 we obtain:

$$\mathbb{E}_{\mathcal{M},s}^{\mathfrak{M}} - \mathbb{E}_{\mathcal{M},s}^{\mathfrak{T} \uparrow r} \geq (\mathbb{CE}^{\text{ub}} - \wp) \cdot (y_s - z_{s,r}) \quad (+)$$

This yields:

$$\begin{aligned} \theta - \zeta &= \sum_{s,r} p_{s,r} \cdot (\mathbb{E}_{\mathcal{M},s}^{\mathfrak{M}} - \mathbb{E}_{\mathcal{M},s}^{\mathfrak{T} \uparrow r}) + \sum_{s,r} p_{s,r} \cdot r \cdot (y_s - z_{s,r}) \\ &\geq (\mathbb{CE}^{\text{ub}} - \wp) \cdot \sum_{s,r} p_{s,r} \cdot (y_s - z_{s,r}) + \sum_{s,r} p_{s,r} \cdot r \cdot (y_s - z_{s,r}) \\ &= (\mathbb{CE}^{\text{ub}} - \wp) \cdot (y - z) + \sum_{s,r} p_{s,r} \cdot r \cdot (y_s - z_{s,r}) \end{aligned}$$

where the sum ranges over all pairs  $(s, r)$  with  $s \in S$  and  $r \in \{\wp, \dots, N\}$ . As

$$\sum_{s \in S} \sum_{r=\wp}^N p_{s,r} \cdot r \cdot (y_s - z_{s,r}) \geq \wp \cdot \sum_{s \in S} \sum_{r=\wp}^N p_{s,r} \cdot (y_s - z_{s,r}) = \wp \cdot (y - z)$$

we obtain:

$$\theta - \zeta \geq (\mathbb{CE}^{\text{ub}} - \wp) \cdot (y - z) + \wp \cdot (y - z) = \mathbb{CE}^{\text{ub}} \cdot (y - z)$$

As  $y_s \geq z_{s,r}$  we have  $y \geq z$ . Let us first consider the case  $y = z$ . By the choice of  $\mathfrak{M}$ ,  $y = z$  implies  $y_s = z_{s,r}$  for all  $s, r$  and therefore  $\mathbb{E}_{\mathcal{M},s}^{\mathfrak{M}} \geq \mathbb{E}_{\mathcal{M},s}^{\mathfrak{T} \uparrow r}$  (see (+)). Consequently,  $\theta \geq \zeta$  and thus  $\mathbb{CE}^{\mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}$ . If  $y > z$  then we get:

$$\frac{\theta - \zeta}{y - z} \geq \mathbb{CE}^{\text{ub}} \geq \mathbb{CE}^{\max} \geq \mathbb{CE}^{\mathfrak{T}}$$

But then  $\mathbb{CE}^{\mathfrak{T}} \leq \mathbb{CE}^{\mathfrak{T} \triangleleft_{\wp} \mathfrak{M}}$  by Lemma E.3. ■

As the cost of computing  $\wp$  is dominated by the computation of  $\mathbb{CE}^{\text{ub}}$ , which has a pseudo-polynomial time bound in the size of  $\mathcal{M}$  (see Appendices C.2 and C.4), we obtain a pseudo-polynomial time bound for the computation of  $\wp$  as well. As the logarithmic length of  $\mathbb{CE}^{\text{ub}}$  is polynomially bounded in the size of  $\mathcal{M}$ , so is the logarithmic length of  $\wp$ .

## G Threshold algorithm

The algorithms for the threshold problem as well as the algorithm to compute the maximal conditional expectation will rely on the following simple observation (Lemma G.1). Among others, we will use it as a semi-local criterion on the best choice for a given state reward pair  $(s, r)$  among two options, say schedulers  $\mathfrak{T}$  and  $\mathfrak{U}$ . Let  $y = \Pr_s^{\mathfrak{T}}(\diamond \text{goal})$  and  $\theta = E_s^{\mathfrak{T}}$ . The pair  $(z, \zeta)$  has analogous meaning for scheduler  $\mathfrak{U}$  where we suppose  $y > z$ . As illustrated by Example 1.1, the best choice might depend on still unknown decisions for other state-reward pairs. In Lemma G.1 these unknown decisions are represented by the parameters  $x, \rho, p$  where  $x$  stands for the probability to reach *goal* from  $s_{\text{init}}$  via path that has no prefix  $\pi$  with  $\text{rew}(\pi) = r$  and  $\text{last}(\pi) = s$  and  $\rho$  for the corresponding expectation. The meaning of  $p$  is the probability to reach  $s$  from  $s_{\text{init}}$  via a path  $\pi$  with  $\text{rew}(\pi) = r$ . Then, Lemma G.1 states that  $\vartheta = r + (\theta - \zeta)/(y - z)$  is a threshold for the decision which of the schedulers  $\mathfrak{T}$  or  $\mathfrak{U}$  is better for  $(s, r)$ :  $\mathfrak{T}$  is better than  $\mathfrak{U}$  if  $\mathbb{CE}^{\text{max}} < \vartheta$  and  $\mathfrak{U}$  is better than  $\mathfrak{T}$  if  $\mathbb{CE}^{\text{max}} > \vartheta$ . Note that given  $\mathfrak{T}$  and  $\mathfrak{U}$ , we can compute  $\vartheta$ , while  $\mathbb{CE}^{\text{max}}$  might still be unknown.

**Lemma G.1.** *Let  $\rho, \theta, \zeta, r, x, y, z, p$  be real numbers such that  $p > 0$ ,  $x, y, z \geq 0$  and  $x+y > 0$  and  $x+z > 0$ . If  $y > z$  then one of the following three cases holds:*

$$\begin{aligned} r + \frac{\theta - \zeta}{y - z} &> \frac{\rho + p(ry + \theta)}{x + py} &> \frac{\rho + p(rz + \zeta)}{x + pz} \\ \text{or} \quad r + \frac{\theta - \zeta}{y - z} &< \frac{\rho + p(ry + \theta)}{x + py} &< \frac{\rho + p(rz + \zeta)}{x + pz} \\ \text{or} \quad r + \frac{\theta - \zeta}{y - z} &= \frac{\rho + p(ry + \theta)}{x + py} &= \frac{\rho + p(rz + \zeta)}{x + pz} \end{aligned}$$

*Proof.* immediate by Lemma E.3. ■

In what follows, we often use Lemma G.1 in the following form. If  $y > z$  then:

$$\begin{aligned} \frac{\rho + p(ry + \theta)}{x + py} &> \frac{\rho + p(rz + \zeta)}{x + pz} &\text{iff} \quad r + \frac{\theta - \zeta}{y - z} &> \frac{\rho + p(ry + \theta)}{x + py} \\ &&&\text{iff} \quad r + \frac{\theta - \zeta}{y - z} &> \frac{\rho + p(rz + \zeta)}{x + pz} \end{aligned}$$

and the analogous statement for  $\geq$  rather than  $>$ .

### G.1 Algorithms for the threshold problem

The input of the threshold problem for maximal conditional expectations is a positive rational number  $\vartheta$  (called threshold) and an MDP  $\mathcal{M}$  with non-negative integer rewards and distinguished states  $s_{init}$ ,  $goal$  and  $fail$ . The goal is to check whether the maximal conditional expectation in  $\mathcal{M}$  meets the bound specified by the threshold value  $\vartheta$  either as a strict or non-strict lower or strict or non-strict upper bound:

$$\begin{aligned} &\text{does } \mathbb{CE}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal | \Diamond goal) \geq \vartheta \text{ hold ?} \\ &\text{does } \mathbb{CE}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal | \Diamond goal) > \vartheta \text{ hold ?} \\ &\text{does } \mathbb{CE}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal | \Diamond goal) \leq \vartheta \text{ hold ?} \\ &\text{does } \mathbb{CE}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal | \Diamond goal) < \vartheta \text{ hold ?} \end{aligned}$$

Throughout this section, we suppose that  $\mathcal{M}$  satisfies conditions (A1), (A2) and has no critical schedulers. Thus,  $\mathbb{CE}^{\max}$  is finite (see Proposition C.8). As before, we write  $\mathbb{CE}^{\max}$  rather than  $\mathbb{CE}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal | \Diamond goal)$ .

Obviously, the threshold problem for strict (resp. non-strict) upper thresholds is dual to the threshold problem for non-strict (resp. strict) thresholds. Thus, it suffices to consider lower thresholds.

As described in Section 4, we provide a threshold algorithm that, given an MDP  $\mathcal{M}$  and a rational threshold  $\vartheta$ , generates a deterministic reward-based scheduler  $\mathfrak{S}$  with  $\mathfrak{S} \uparrow \varphi = \mathfrak{M}$  (where  $\mathfrak{M}$  and  $\varphi$  are as in Prop. 4.1) such that  $\mathbb{CE}^{\mathfrak{S}} > \vartheta$  if  $\mathbb{CE}^{\max} > \vartheta$ , and  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$  if  $\mathbb{CE}^{\max} = \vartheta$ . If  $\mathbb{CE}^{\max} < \vartheta$  then the output of the threshold algorithm is “no”. It is easy to see how this algorithm can be used in a decision procedure for deciding whether  $\mathbb{CE}^{\mathfrak{S}} > \vartheta$  or  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$ .

In a preprocessing step, we compute the saturation point  $\varphi$  (see Section F). The threshold algorithm operates level-wise and attempts to construct a reward-based deterministic scheduler that is memoryless from the last level  $\varphi$  and that satisfies the threshold condition  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$ , provided  $\mathbb{CE}^{\max} \geq \vartheta$ . Otherwise the algorithm returns “no”.

*Initialization of the threshold algorithm.* The treatment of level  $\varphi$  is obvious as optimal decisions are known by the results of Appendix E. Let  $\mathfrak{M}$  be a deterministic memoryless scheduler that maximizes the probability to reach  $goal$  from each state and whose conditional expectation is maximal under all those schedulers (by Lemma E.14). Let  $action(s, \varphi) \in Act(s)$  be the action that  $\mathfrak{M}$  chooses for state  $s$ . Furthermore, we define  $y_{s, \varphi} = p_s^{\max}$  and  $\theta_{s, \varphi} = E_s^{\mathfrak{M}}$ .

*Level-wise computation of feasible actions.* For  $r = \varphi - 1, \varphi - 2, \dots, 2, 1, 0$  and each state  $s \in S \setminus \{goal, fail\}$ , the algorithm computes actions  $action(s, r) \in Act(s)$  and rational values  $y_{s, r}$ ,  $\theta_{s, r}$  for the probability to reach  $goal$  from  $s$  and the corresponding expectation under the residual scheduler defined by the action table  $action(\cdot)$ . The values for the trap states are the trivial ones:  $y_{goal, r} = 1$ ,  $y_{fail, r} = 0$  and  $\theta_{goal, r} = \theta_{fail, r} = 0$ .

Suppose now that  $r \in \mathbb{N}$  with  $0 \leq r < \varphi$  and that levels  $r+1, r+2, \dots, \varphi$  have been treated before and the triples  $(action(t, R), y_{t, R}, \theta_{t, R})$  have been computed

for all states  $t \in S$  and for all  $R \in \{r+1, \dots, \wp\}$ . Before describing the steps that the threshold algorithm performs to treat level  $r$ , we explain the demands on the actions that the threshold algorithm assigns to the state-reward pairs  $(s, r)$ .

*Most feasible actions at level  $r$ .* The goal of the treatment of level  $r$  is to find the most feasible way to combine zero-reward actions with positive-reward actions as decisions of a deterministic reward-based scheduler for paths where the accumulated reward is  $r$ . Here, “most feasible” is understood with respect to the task to assign actions to the state-reward pairs  $(s, r)$  that any scheduler whose conditional expectation is at least or larger than the given threshold  $\vartheta$  may take. More precisely, given a function  $\mathfrak{P} : S \setminus \{goal, fail\} \rightarrow Act$  such that  $\mathfrak{P}(s) \in Act(s)$  for all states  $s$ , let

$$T_{\mathfrak{P}} = \{goal, fail\} \cup \{s \in S \setminus \{goal, fail\} : rew(s, \mathfrak{P}(s)) > 0\}$$

For the non-trap states  $s \in T_{\mathfrak{P}} \setminus \{goal, fail\}$ , the values  $y_{s,r,\mathfrak{P}}$  and  $\theta_{s,r,\mathfrak{P}}$  are defined by:

$$\begin{aligned} y_{s,r,\mathfrak{P}} &= \sum_{t \in S} P(s, \mathfrak{P}(s), t) \cdot y_{t,R} \\ \theta_{s,r,\mathfrak{P}} &= rew(s, \alpha) \cdot y_{s,r,\mathfrak{P}} + \sum_{t \in S} P(s, \mathfrak{P}(s), t) \cdot \theta_{t,R} \end{aligned}$$

where  $R = \min\{\wp, r + rew(s, \alpha)\}$ . Furthermore,  $y_{fail,r,\mathfrak{P}} = \theta_{fail,r,\mathfrak{P}} = \theta_{goal,r,\mathfrak{P}} = 0$  and  $y_{goal,r,\mathfrak{P}} = 1$ . For the states  $s \in S \setminus T_{\mathfrak{P}}$  we define:

$$\begin{aligned} y_{s,r,\mathfrak{P}} &= \sum_{t \in T_{\mathfrak{P}}} \Pr_s^{\mathfrak{P}}(\neg T_{\mathfrak{P}} \cup t) \cdot y_{t,r,\mathfrak{P}} \\ \theta_{s,r,\mathfrak{P}} &= \sum_{t \in T_{\mathfrak{P}}} \Pr_s^{\mathfrak{P}}(\neg T_{\mathfrak{P}} \cup t) \cdot \theta_{t,r,\mathfrak{P}} \end{aligned}$$

The task of the threshold algorithm is now to find a function  $\mathfrak{P}^*$  satisfying the following constraint. If  $\mathbb{CE}^{\max} \geq \vartheta$  then there exists an eventually memoryless, reward-based scheduler  $\mathfrak{T}$  with  $\Pr_{s_{init}}^{\mathfrak{T}}(\Diamond goal) > 0$  and  $\mathbb{CE}^{\mathfrak{T}} \geq \vartheta$  that schedules

- $\mathfrak{M}(s)$  for all state-reward pairs  $(s, R)$  with  $R \geq \wp$  and
- the actions  $action(t, R)$  for all state-reward pairs  $(t, R)$  with  $r < R < \wp$
- $\mathfrak{P}^*(s)$  for the state-reward pairs  $(s, r)$ .

To do so, we present a procedure that is based on linear programming techniques and the following observation.

**Lemma G.2.** *Let  $\rho, \theta, \zeta, \vartheta, p, x, y, z$  be rational numbers with  $p > 0$ ,  $x, y, z \geq 0$ ,  $x+y, x+z > 0$  and  $r \in \mathbb{N}$  such that*

$$\theta - (\vartheta - r) \cdot y \geq \zeta - (\vartheta - r) \cdot z \quad (*)$$

Then:

$$\begin{aligned}
 \text{(a)} \quad & \frac{\rho + p(rz + \zeta)}{x + pz} \geq \vartheta \quad \text{implies} \quad \frac{\rho + p(ry + \theta)}{x + py} \geq \vartheta \\
 \text{(b)} \quad & \frac{\rho + p(rz + \zeta)}{x + pz} > \vartheta \quad \text{implies} \quad \frac{\rho + p(ry + \theta)}{x + py} > \vartheta
 \end{aligned}$$

Before presenting the proof of Lemma G.2, let us first state its role for the treatment of level  $r$  in the threshold algorithm. Intuitively, the value  $x$  stands for the probability to reach *goal* along some path that has no prefix  $\pi$  with  $\text{rew}(\pi) = r$  and  $\rho$  for the corresponding expectation, while  $p$  denotes the probability to generate a finite path  $\pi$  from  $s_{\text{init}}$  to some state  $s$  with  $\text{rew}(\pi) = r$ . The pairs  $(\theta, y)$  and  $(\zeta, z)$  stand for the expected reward and probability to reach *goal* from  $s$  under some scheduler. When treating the states at level  $r$ , the values  $\rho, x, p$  are unknown, while the pairs  $(\theta_{s,r,\mathfrak{P}}, y_{s,r,\mathfrak{P}})$  are candidates for  $(\theta, y)$  and  $(\zeta, z)$ . Thus, Lemma G.2 asserts that the most promising candidates for  $\mathfrak{P}$  are the ones where  $\theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}}$  is maximal for all states  $s$ . Here, “most promising” means that if  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$  for some scheduler  $\mathfrak{S}$  that extends the already made decisions for levels  $r+1, \dots, \wp$  then  $\mathbb{CE}^{\mathfrak{T}} \geq \vartheta$  for some scheduler  $\mathfrak{S}$  that extends the decisions for levels  $r+1, \dots, \wp$  and behaves as  $\mathfrak{P}$  for level  $r$ .

*Proof.* We first consider statement (a) and suppose  $(\rho + p(rz + \zeta))/(x + pz) \geq \vartheta$ . The task is to show that (\*) implies  $(\rho + p(ry + \theta))/(x + py) \geq \vartheta$ . The claim is clear if  $y = z$ , in which case (\*) implies  $\theta \geq \zeta$ . Suppose now  $y > z$ . Then, (\*) implies

$$\frac{\theta - \zeta}{y - z} \geq \vartheta - r$$

and therefore

$$r + \frac{\theta - \zeta}{y - z} \geq \vartheta$$

Suppose by contradiction that

$$\frac{\rho + p(ry + \theta)}{x + py} < \vartheta$$

Then:

$$\frac{\rho + p(ry + \theta)}{x + py} < r + \frac{\theta - \zeta}{y - z}$$

We now apply Lemma G.1 and obtain:

$$\frac{\rho + p(ry + \theta)}{x + py} > \frac{\rho + p(rz + \zeta)}{x + pz} \geq \vartheta$$

Contradiction. Hence,  $(\rho + p(ry + \theta))/(x + py) \geq \vartheta$  if  $y \geq z$ . The remaining case  $y < z$  can be handled by analogous arguments. This completes the proof of Lemma G.2.  $\blacksquare$

Minimize  $\sum_{s \in S} x_s$  subject to:

(1) If  $s \in S \setminus \{goal, fail\}$  then for each action  $\alpha \in Act(s)$  with  $rew(s, \alpha) = 0$ :

$$x_s \geq \sum_{t \in S} P(s, \alpha, t) \cdot x_t$$

(2) If  $s \in S \setminus \{goal, fail\}$  then for each action  $\alpha \in Act(s)$  with  $rew(s, \alpha) > 0$ :

$$x_s \geq \sum_{t \in S} P(s, \alpha, t) \cdot (\theta_{t,R} + rew(s, \alpha) \cdot y_{t,R} - (\vartheta - r) \cdot y_{t,R})$$

where  $R = \min\{\wp, r + rew(s, \alpha)\}$

(3) For the trap states:  $x_{goal} = r - \vartheta$  and  $x_{fail} = 0$

**Fig. 3.** Linear program for the treatment of level  $r$  in the threshold algorithm

The idea for the treatment of each level  $r < \wp$  is now to compute the values  $\max_{\mathfrak{P}}(\theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}})$  for all states by solving the linear program shown in Figure 3. The latter has one variable  $x_s$  for each state  $s \in S$  and one linear constraint for each state-action pair  $(s, \alpha)$  with  $\alpha \in Act(s)$ .

Lemma G.3 (see below) will show the existence of a unique solution of the linear program in Figure 3. Let  $(x_s^*)_{s \in S}$  be the solution of the linear program in Figure 3. Let  $Act^*(s)$  denote the set of actions  $\alpha \in Act(s)$  such that the following constraints (E1) and (E2) hold:

$$(E1) \quad \text{If } rew(s, \alpha) = 0 \text{ then: } x_s^* = \sum_{t \in S} P(s, \alpha, t) \cdot x_t^*$$

$$(E2) \quad \text{If } rew(s, \alpha) > 0 \text{ and } R = \min\{\wp, r + rew(s, \alpha)\} \text{ then:}$$

$$x_s^* = \sum_{t \in S} P(s, \alpha, t) \cdot (\theta_{t,R} + rew(s, \alpha) \cdot y_{t,R} - (\vartheta - r) \cdot y_{t,R})$$

Let  $\mathcal{M}^* = \mathcal{M}_{r,\vartheta}^*$  denote the MDP with state space  $S$  induced by the state-action pairs  $(s, \alpha)$  with  $\alpha \in Act^*(s)$  where the positive-reward actions are redirected to the trap states. More precisely, if  $s, t \in S$  and  $\alpha \in Act^*(s)$  and  $rew(s, \alpha) = 0$  then  $P_{\mathcal{M}^*}(s, \alpha, t) = P(s, \alpha, t)$ . For  $\alpha \in Act^*(s)$  and  $rew(s, \alpha) > 0$ :

$$P_{\mathcal{M}^*}(s, \alpha, goal) = \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,R}, \quad P_{\mathcal{M}^*}(s, \alpha, fail) = 1 - P_{\mathcal{M}^*}(s, \alpha, goal)$$

where  $R = \min\{\wp, r + rew(s, \alpha)\}$ . The reward structure of  $\mathcal{M}^*$  is irrelevant for our purposes.

*Treatment of level  $r$  in the threshold algorithm.* The threshold algorithm solves the linear program of Figure 3<sup>11</sup> and then computes the action sets  $Act^*(s)$  and a deterministic memoryless scheduler  $\mathfrak{P}^* : S \setminus \{goal, fail\} \rightarrow Act$  for the MDP  $\mathcal{M}^*$  with

<sup>11</sup> Note that the values  $\theta_{t,R}$  and  $y_{t,R}$  for  $t \in S$  and  $r < R \leq \wp$  used in the constraints (2) of Figure 3 have been computed before in the treatment of level  $R$ .

$$\mathfrak{P}^*(s) \in Act^*(s) \quad \text{and} \quad \Pr_{\mathcal{M}^*,s}^{\mathfrak{P}^*}(\Diamond goal) = \Pr_{\mathcal{M}^*,s}^{\max}(\Diamond goal)$$

for all  $s \in S \setminus \{goal, fail\}$ . It then defines:

$$action(s, r) = \mathfrak{P}^*(s), \quad y_{s,r} = y_{s,r,\mathfrak{P}^*} \quad \text{and} \quad \theta_{s,r} = \theta_{s,r,\mathfrak{P}^*}.$$

This completes the treatment of level  $r$ .

*Output of the threshold algorithm.* Having reached the last level  $r = 0$ , the output of the algorithm is as follows. The generated reward-based scheduler  $\mathfrak{S}$ , given by  $\mathfrak{S}(s, r) = action(s, r)$  for  $r < \wp$  and  $\mathfrak{S}(s, r) = \mathfrak{M}(s)$  for  $r \geq \wp$ , satisfies the equations  $y_{s_{init},0} = \Pr_{s_{init}}^{\mathfrak{S}}(\Diamond goal)$  and  $\theta_{s_{init},0} = E_{s_{init}}^{\mathfrak{S}}$ . Thus, if  $y_{s_{init},0} > 0$  then  $\mathbb{CE}^{\mathfrak{S}} = \theta_{s_{init},0}/y_{s_{init},0}$ . Hence, the threshold algorithm returns  $\mathfrak{S}$  if  $y_{s_{init},0} > 0$  and  $\theta_{s_{init},0}/y_{s_{init},0} \geq \vartheta$ . The correctness is obvious, as  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$ . Otherwise, i.e., if  $y_{s_{init},0} = 0$  or  $\theta_{s_{init},0}/y_{s_{init},0} < \vartheta$ , the algorithm terminates with the answer “no”. The correctness of the answer “no” is a consequence of Lemma G.4 (see below).

**Lemma G.3 (Soundness of the LP of Figure 3).** *The linear program in Figure 3 has a unique solution  $(x_s^*)_{s \in S}$ . The action-sets  $Act^*(s)$  are non-empty for all  $s \in S \setminus \{goal, fail\}$  and for each function  $\mathfrak{P} : S \setminus \{goal, fail\} \rightarrow Act$  with  $\mathfrak{P}(s) \in Act^*(s)$  we have:*

$$x_s^* = \theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}}$$

Moreover, whenever  $\mathfrak{P} : S \setminus \{goal, fail\} \rightarrow Act$  is a function with  $\mathfrak{P}(s) \in Act(s)$  then:

$$x_s^* \geq \theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}}$$

*Proof.* The solvability and the uniqueness of the solution of the linear program in Figure 3 follows by the fact that the linear program agrees with the one that is known to represent the expected total reward in the MDP  $\mathcal{N}$  with state space  $S_{\mathcal{N}} = S \cup \{final\}$  and action set  $Act_{\mathcal{N}} = Act \cup \{\tau\}$  such that for all states  $s \in S$  and all actions  $\alpha \in Act(s)$ :

$$P_{\mathcal{N}}(s, \alpha, t) = P(s, \alpha, t) \quad \text{and} \quad rew_{\mathcal{N}}(s, \alpha) = 0 \quad \text{if} \quad rew(s, \alpha) = 0$$

$$P_{\mathcal{N}}(s, \alpha, final) = 1 \quad \text{and} \quad rew_{\mathcal{N}}(s, \alpha) = \theta_{s,r,\alpha} - (\vartheta - r) \cdot y_{s,r,\alpha} \quad \text{if} \quad rew(s, \alpha) > 0$$

$$P_{\mathcal{N}}(goal, \tau, final) = 1 \quad \text{and} \quad rew_{\mathcal{N}}(goal, \tau) = r - \vartheta$$

$$P_{\mathcal{N}}(fail, \tau, final) = 1 \quad \text{and} \quad rew_{\mathcal{N}}(fail, \tau) = 0$$

and  $P_{\mathcal{N}}(\cdot) = rew_{\mathcal{N}}(\cdot) = 0$  in all remaining cases. In particular, state *final* is a trap in  $\mathcal{N}$  and there are no other traps in  $\mathcal{N}$ . Assumption (A2) yields  $\Pr_{\mathcal{N},s}^{\min}(\Diamond final) = 1$  for all states  $s \in S_{\mathcal{N}}$ . Using standard results for finite-state MDPs (see e.g. Theorem 4.20 in [28]), the values  $E_{\mathcal{N},s}^{\max}(\Diamond final)$  are finite and are computable as the unique solution of the linear program shown in Figure 3. That is, if  $(x_s^*)_{s \in S}$  is the unique solution of the linear program in Figure 3 then

$$x_s^* = E_{\mathcal{N},s}^{\max}(\Diamond final) \quad \text{for all states } s \in S.$$

Let now  $\mathfrak{P}$  be a deterministic memoryless scheduler for  $\mathcal{N}$  that maximizes the expected total reward in  $\mathcal{N}$ , i.e.,  $x_s^* = E_{\mathcal{N},s}^{\mathfrak{P}}(\Diamond final)$  for all states  $s \in S$ . Clearly,  $\mathfrak{P}$  can be viewed as a function  $S \setminus \{goal, fail\} \rightarrow Act$  with  $\mathfrak{P}(s) \in Act^*(s)$  for all  $s \in S$ . This yields that the sets  $Act^*(s)$  are non-empty for the non-trap states  $s$  for  $\mathcal{M}$ . Vice versa, each function  $\mathfrak{P} : S \setminus \{goal, fail\} \rightarrow Act$  with  $\mathfrak{P}(s) \in Act^*(s)$  for all  $s \in S \setminus \{goal, fail\}$  can be viewed as a deterministic memoryless scheduler for  $\mathcal{N}$  that maximizes the expected total reward in  $\mathcal{N}$ .

Suppose now that  $\mathfrak{P} : S \setminus \{goal, fail\} \rightarrow Act$  is a function with  $\mathfrak{P}(s) \in Act(s)$  for all  $s \in S \setminus \{goal, fail\}$ . Let  $T$  denote the set of states  $s \in S$  such that either  $s \in \{goal, fail\}$  or  $rew(s, \mathfrak{P}(s)) > 0$ . Clearly,  $\mathfrak{P}$  can be viewed as a scheduler for  $\mathcal{N}$  that schedules the unique action  $\tau$  for the trap-states  $goal$  and  $fail$  of  $\mathcal{M}$ , and we have:

$$\Pr_{\mathcal{N},s}^{\mathfrak{P}}(\Diamond T) = 1$$

With  $\theta_{fail,r,\mathfrak{P}} = \theta_{fail,r} = 0$ ,  $\theta_{goal,r,\mathfrak{P}} = \theta_{goal,r} = 0$ ,  $y_{fail,r,\mathfrak{P}} = y_{fail,r} = 0$  and  $y_{goal,r,\mathfrak{P}} = y_{goal,r} = 1$  we obtain:

$$rew_{\mathcal{N}}(t, \mathfrak{P}(t)) = \theta_{t,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{t,r,\mathfrak{P}}$$

for all states  $t \in T$ . This yields that for all states  $s \in S$ :

$$\begin{aligned} E_{\mathcal{N},s}^{\mathfrak{P}}(\Diamond final) &= \sum_{t \in T} \Pr_{\mathcal{N},s}^{\mathfrak{P}}(\neg T \cup t) \cdot rew_{\mathcal{N}}(t, \mathfrak{P}(t)) \\ &= \sum_{t \in T} \Pr_{\mathcal{N},s}^{\mathfrak{P}}(\neg T \cup t) \cdot \theta_{t,r,\mathfrak{P}} - (\vartheta - r) \cdot \sum_{t \in T} \Pr_{\mathcal{N},s}^{\mathfrak{P}}(\neg T \cup t) \cdot y_{t,r,\mathfrak{P}} \\ &= \theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}} \end{aligned}$$

As  $x_s^* = E_{\mathcal{N},s}^{\max}(\Diamond final) \geq E_{\mathcal{N},s}^{\mathfrak{P}}(\Diamond final)$  we obtain:

$$x_s^* \geq \theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}}$$

Moreover, if  $\mathfrak{P}(s) \in Act^*(s)$  for all states  $s$  then  $x_s^* = E_{\mathcal{N},s}^{\max}(\Diamond final) = E_{\mathcal{N},s}^{\mathfrak{P}}(\Diamond final)$  (see above) and therefore:

$$x_s^* = E_{\mathcal{N},s}^{\max}(\Diamond final) = E_{\mathcal{N},s}^{\mathfrak{P}}(\Diamond final) = \theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}}$$

This completes the proof of Lemma G.3.  $\blacksquare$

It remains to show that if the algorithm returns “no” then there is no scheduler meeting the bound for its conditional expectation. We prove this by showing that if  $\mathbb{CE}^{\max} \geq \vartheta$  then after the treatment of each level  $r$  there exists a reward-based scheduler  $\mathfrak{T}_r$  using the decisions that have been stored in the action-table  $action(\cdot)$  for all level  $> r$  and the decisions of  $\mathfrak{P}^*$  at level  $r$  and satisfying  $\mathbb{CE}^{\mathfrak{T}_r} \geq \vartheta$ .

**Lemma G.4 (Soundness of the answer “no”).** *If  $\mathbb{CE}^{\max} \geq \vartheta$  then the threshold algorithm generates a scheduler  $\mathfrak{S}$  with  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$ .*



*Proof.* The task is to show that the algorithm indeed returns a scheduler  $\mathfrak{S}$  with  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$  if  $\mathbb{CE}^{\max} \geq \vartheta$ . For this, we use an inductive argument to prove the following statement. If  $\mathbb{CE}^{\max} \geq \vartheta$  then for each  $r \in \{\wp, \wp-1, \dots, 1, 0\}$ , there exists a reward-based scheduler  $\mathfrak{T}_r$  for  $\mathcal{M}$  with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{T}_r}(\Diamond goal) > 0$  and  $\mathbb{CE}^{\mathfrak{T}_r} \geq \vartheta$  such that:

- $\mathfrak{T}_r(s, R) = \mathfrak{M}(s)$  for all state-reward pairs  $(s, R)$  with  $R \geq \wp$  and
- $\mathfrak{T}_r(s, R) = action(t, R)$  for all state-reward pairs  $(t, R)$  with  $r < R < \wp$
- $\mathfrak{T}_r(s, r) = \mathfrak{P}^*(s)$  for the state-reward pairs  $(s, r)$

where  $\mathfrak{P}^*$  is the function as explained in the treatment of level  $r$ . For  $r = 0$  we obtain  $\mathfrak{T}_0 = \mathfrak{S}$  and therefore  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$ .

The claim is obvious for  $r = \wp$  as then we can deal with  $\mathfrak{T}_{\wp} = \mathfrak{T}$  where  $\mathfrak{T}$  is any scheduler with  $\mathbb{CE}^{\mathfrak{T}} = \mathbb{CE}^{\max}$ . (This follows from the fact that  $\wp > \mathfrak{R}$  for the turning point  $\mathfrak{R}$  of Proposition E.8.) Suppose now that  $r < \wp$ . By induction hypothesis there exists a reward-based scheduler  $\mathfrak{T}_{r+1} = \mathfrak{U}$  such that

- $\mathfrak{U}(s, R) = \mathfrak{M}(s)$  for all state-reward pairs  $(s, R)$  with  $R \geq \wp$  and
- $\mathfrak{U}(s, R) = action(t, R)$  for all state-reward pairs  $(t, R)$  with  $r < R < \wp$

and  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{U}}(\Diamond goal) > 0$  and  $\mathbb{CE}^{\mathfrak{U}} \geq \vartheta$ . Let

$$p_s = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{U}}(\Diamond^{=r} s), \quad p = \sum_{s \in S} p_s$$

Furthermore, we define  $\mathfrak{P} : S \setminus \{goal, fail\} \rightarrow Act$  by  $\mathfrak{P}(s) = \mathfrak{U}(s, r)$ . Then:

$$\zeta_s \stackrel{\text{def}}{=} \theta_{s,r,\mathfrak{P}} = E_{\mathcal{M},s}^{\mathfrak{U} \uparrow r}, \quad z_s \stackrel{\text{def}}{=} y_{s,r,\mathfrak{P}} = \Pr_{\mathcal{M},s}^{\mathfrak{U} \uparrow r}(\Diamond goal)$$

Likewise, we define

$$\theta_s = \theta_{s,r,\mathfrak{P}^*} \quad \text{and} \quad y_s = \theta_{s,r,\mathfrak{P}^*}$$

Let  $\mathfrak{T} = \mathfrak{T}_r$  denote the scheduler for  $\mathcal{M}$  that agrees with  $\mathfrak{U}$  except that  $\mathfrak{T}(s, r) = \mathfrak{P}^*(s)$  for all states  $s$  of  $\mathcal{M}$ . We then have

$$\theta_s = E_{\mathcal{M},s}^{\mathfrak{T} \uparrow r}(\Diamond goal) \quad \text{and} \quad y_s = \Pr_{\mathcal{M},s}^{\mathfrak{P}^* \uparrow r}(\Diamond goal).$$

Let

$$\begin{aligned} \theta &= \sum_{s \in S} \frac{p_s}{p} \cdot \theta_s, & y &= \sum_{s \in S} \frac{p_s}{p} \cdot y_s \\ \zeta &= \sum_{s \in S} \frac{p_s}{p} \cdot \zeta_s, & z &= \sum_{s \in S} \frac{p_s}{p} \cdot z_s \end{aligned}$$

Then,  $py$  is the probability of the infinite  $\mathfrak{T}$ -paths from  $s_{init}$  that have a prefix  $\pi$  with  $rew(\pi) = r$  and  $p\theta$  the corresponding expectation. The values  $pz$  and  $p\zeta$  have analogous meaning for scheduler  $\mathfrak{U}$ .

We now use Lemma G.3. As  $x_s^* = \theta_{s,r,\mathfrak{P}^*} - (\vartheta-r) \cdot y_{s,r,\mathfrak{P}^*}$  and  $x_s^* \geq \theta_{s,r,\mathfrak{P}} - (\vartheta-r) \cdot y_{s,r,\mathfrak{P}}$ , we have:

$$\begin{aligned} \theta_s - (\vartheta-r) \cdot y_s &= \theta_{s,r,\mathfrak{P}^*} - (\vartheta-r) \cdot y_{s,r,\mathfrak{P}^*} \\ &\geq \theta_{s,r,\mathfrak{P}} - (\vartheta-r) \cdot y_{s,r,\mathfrak{P}} = \zeta_s - (\vartheta-r) \cdot z_s \end{aligned}$$

for all states  $s \in S$ . But this yields:

$$\theta - (\vartheta-r) \cdot y \geq \zeta - (\vartheta-r) \cdot z$$

There exists non-negative rational numbers  $\rho$  and  $x$  (namely,  $x$  is the probability of the maximal  $\mathfrak{U}$ -paths from  $s_{init}$  to  $goal$  that do not have a prefix  $\pi$  with  $rew(\pi) = r$  and  $\rho$  is the corresponding expectation) such that

$$\mathbb{CE}^{\mathfrak{U}} = \frac{\rho + p(rz + \zeta)}{x + pz} \quad \text{and} \quad \mathbb{CE}^{\mathfrak{T}} = \frac{\rho + p(ry + \theta)}{x + py}$$

Let us check that  $x + py$  is indeed positive. This is clear if  $x > 0$ . Suppose now that  $x = 0$ . The goal is to show that  $y > 0$ . As  $x = 0$  we have  $\rho = 0$  and therefore:

$$\vartheta \leq \mathbb{CE}^{\mathfrak{U}} = \frac{p(rz + \zeta)}{pz} = r + \frac{\zeta}{z}$$

Hence,  $(\vartheta-r) \cdot z \leq \zeta$ . Suppose by contradiction that  $y = 0$ . Then,  $\theta = 0$ . Hence,

$$0 = \theta - (\vartheta-r)y \geq \zeta - (\vartheta-r)z \geq 0$$

This yields  $0 = \theta - (\vartheta-r)y = \zeta - (\vartheta-r)z$ . As  $\theta_s - (\vartheta-r)y_s \geq \zeta_s - (\vartheta-r)z_s$  for all states  $s$ , we have  $\theta_s - (\vartheta-r)y_s = \zeta_s - (\vartheta-r)z_s$ . This implies that  $\mathfrak{P}$  viewed as a scheduler for  $\mathcal{N}$  maximizes the expected total reward from every state  $s$ . In particular,  $\mathfrak{P}(s) \in Act^*(s)$  for all states  $s$ . Thus,  $\mathfrak{P}$  can also be viewed as a scheduler for the MDP  $\mathcal{M}^*$ . As  $x + pz > 0$  and  $x = 0$  (by assumption) we have  $z > 0$ . Thus,  $z_s > 0$  for at least one state  $s$  with  $p_s > 0$ . By the choice of  $\mathfrak{P}^*$ ,

$$y_s = \Pr_{\mathcal{M}^*,s}^{\mathfrak{P}^*}(\Diamond goal) = \Pr_{\mathcal{M}^*,s}^{\max}(\Diamond goal) \geq \Pr_{\mathcal{M}^*,s}^{\mathfrak{P}}(\Diamond goal) = z_s > 0$$

This yields  $y > 0$ , and therefore  $x + py > 0$ .

We are now in the position to apply Lemma G.2. Recall that we have

$$\vartheta \leq \mathbb{CE}^{\mathfrak{U}} = \frac{\rho + p(rz + \zeta)}{x + pz} \quad \text{and} \quad \mathbb{CE}^{\mathfrak{T}} = \frac{\rho + p(ry + \theta)}{x + py}$$

and

$$\theta - (\vartheta-r) \cdot y \geq \zeta - (\vartheta-r) \cdot z$$

Part (a) of Lemma G.2 yields  $\mathbb{CE}^{\mathfrak{T}} \geq \vartheta$ . ■

**Corollary G.5 (Optimality of the generated scheduler).** *If the threshold algorithm returns a scheduler  $\mathfrak{S}$  with  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$  then  $\mathbb{CE}^{\max} = \vartheta$  and  $\mathfrak{S}$  is a reward-based scheduler that maximizes the conditional expectation.*

The above corollary as well as the following observations about the scheduler  $\mathfrak{S}$  generated by the threshold algorithm will be crucial for the computation of the maximal conditional expectation  $\mathbb{CE}^{\max}$ .

Lemma G.3 yields for the function  $\mathfrak{P}^*$  used to define the decisions of the generated scheduler at level  $r$ :

$$\theta_{s,r,\mathfrak{P}^*} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}^*} = \max_{\mathfrak{P}} (\theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}})$$

where  $\mathfrak{P}$  ranges over all functions  $\mathfrak{P} : S \setminus \{\text{goal}, \text{fail}\} \rightarrow \text{Act}$  with  $\mathfrak{P}(s) \in \text{Act}(s)$  for all states  $s \in S$ . This implies the first part of the following lemma:

**Lemma G.6 (Difference-property of the generated scheduler).** *Notations as before. The scheduler  $\mathfrak{S}$  generated by the threshold algorithm for the threshold  $\vartheta$  enjoys the following property. For each  $r \in \{0, 1, \dots, \wp\}$ , each state  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and each function  $\mathfrak{P} : S \setminus \{\text{goal}, \text{fail}\} \rightarrow \text{Act}$  with  $\mathfrak{P}(t) \in \text{Act}(t)$  for all  $t$  we have:*

$$\theta_{s,r} - (\vartheta - r) \cdot y_{s,r} \geq \theta_{s,r,\mathfrak{P}} - (\vartheta - r) \cdot y_{s,r,\mathfrak{P}}$$

Moreover, if  $r \in \{0, 1, \dots, \wp\}$  and

$$\vartheta_r = \min \left\{ r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} : s \in S \setminus \{\text{goal}, \text{fail}\}, y_{s,r} > y_{s,r,\alpha} \right\}$$

(where we put  $\min \emptyset = +\infty$ ) then  $\vartheta_r \geq \vartheta$  and for each value  $\vartheta^*$  with  $\vartheta^* \geq \vartheta$  we have:  $\vartheta^*$  satisfies the following condition (\*) for all states  $s$  and functions  $\mathfrak{P}$  if and only if  $\vartheta^* \leq \vartheta_r$ .

$$\theta_{s,r} - (\vartheta^* - r) \cdot y_{s,r} \geq \theta_{s,r,\mathfrak{P}} - (\vartheta^* - r) \cdot y_{s,r,\mathfrak{P}} \quad (*)$$

Furthermore, if  $\vartheta < \vartheta^* \leq \min\{\vartheta_R : r \leq R \leq \wp\}$  and  $\mathfrak{S}^*$  is the scheduler that has been generated by the threshold algorithm for the lower bound  $\vartheta^*$  then  $(y_{s,R}, \theta_{s,R}) = (y_{s,R}^*, \theta_{s,R}^*)$  for all states  $s \in S$  and  $R \in \{r, \dots, \wp\}$  where

$$y_{s,R}^* = \Pr_s^{\mathfrak{S}^* \uparrow r}(\Diamond \text{goal}), \quad \theta_{s,R}^* = E_s^{\mathfrak{S}^* \uparrow r}$$

These properties hold, no matter whether  $\mathbb{CE}^{\mathfrak{S}} < \vartheta$  or  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$  or  $\mathbb{CE}^{\mathfrak{S}} > \vartheta$ .

*Proof.* The first statement is obvious. To prove the second statement, we pick some  $r \in \{0, 1, \dots, \wp\}$ . Obviously, we have  $\vartheta_r \geq \vartheta$ .

Recall that  $y_{s,r,\alpha} = \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,R}$  and  $\theta_{s,r,\alpha} = \sum_{t \in S} P(s, \alpha, t) \cdot \theta_{t,R}$  where  $R = \min\{\wp, r + \text{rew}(s, \alpha)\}$ .

Let now  $\vartheta^*$  be any value with  $\vartheta \leq \vartheta^* \leq \vartheta_r$ . We prove that (\*) holds. As  $\vartheta \leq \vartheta^*$  we obtain:

$$\theta_{s,r} - (\vartheta^* - r) \cdot y_{s,r} \geq \theta_{s,r,\alpha} - (\vartheta^* - r) \cdot y_{s,r,\alpha}$$

for all states  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and actions  $\alpha \in \text{Act}(s)$ .

As in the proof of Lemma G.3, we consider the MDP  $\mathcal{N}$  with state space  $S_{\mathcal{N}} = S \cup \{\text{final}\}$  and action set  $\text{Act}_{\mathcal{N}} = \text{Act} \cup \{\tau\}$  such that for all states  $s \in S$  and all actions  $\alpha \in \text{Act}(s)$ :

$$\begin{aligned}
P_{\mathcal{N}}(s, \alpha, t) &= P(s, \alpha, t) \text{ and } \text{rew}_{\mathcal{N}}(s, \alpha) = 0 \quad \text{if } \text{rew}(s, \alpha) = 0 \\
P_{\mathcal{N}}(s, \alpha, \text{final}) &= 1 \text{ and } \text{rew}_{\mathcal{N}}(s, \alpha) = \theta_{s, \alpha} - (\vartheta^* - r) \cdot y_{s, \alpha} \quad \text{if } \text{rew}(s, \alpha) > 0 \\
P_{\mathcal{N}}(\text{goal}, \tau, \text{final}) &= 1 \text{ and } \text{rew}_{\mathcal{N}}(\text{goal}, \tau) = r - \vartheta^* \\
P_{\mathcal{N}}(\text{fail}, \tau, \text{final}) &= 1 \text{ and } \text{rew}_{\mathcal{N}}(\text{fail}, \tau) = 0
\end{aligned}$$

and  $P_{\mathcal{N}}(\cdot) = \text{rew}_{\mathcal{N}}(\cdot) = 0$  in all remaining cases. Note that state *final* is a trap in  $\mathcal{N}$  and there are no other traps in  $\mathcal{N}$ . Assumption (A2) yields  $\Pr_{\mathcal{N}, s}^{\min}(\Diamond \text{final}) = 1$  for all states  $s \in S_{\mathcal{N}}$ . For  $s \in S$  let

$$x_s^{\mathfrak{S}} = \theta_{s, r} - (\vartheta^* - r) \cdot y_{s, r}$$

and let  $x_{\text{final}}^{\mathfrak{S}} = 0$ . For  $\alpha \in \text{Act}(s)$  with  $\text{rew}(s, \alpha) > 0$  and  $R = \min\{\wp, r + \text{rew}(s, \alpha)\}$  we have:

$$\begin{aligned}
x_s^{\mathfrak{S}} &\geq \theta_{s, r, \alpha} - (\vartheta^* - r) \cdot y_{s, r, \alpha} \\
&= \text{rew}_{\mathcal{N}}(s, \alpha) + P_{\mathcal{N}}(s, \alpha, \text{final}) \cdot x_{\text{final}}^{\mathfrak{S}} \\
&= \text{rew}_{\mathcal{N}}(s, \alpha) + \sum_{t \in S_{\mathcal{N}}} P(s, \alpha, t) \cdot x_t^{\mathfrak{S}}
\end{aligned}$$

Thus, the vector  $(x_s^{\mathfrak{S}})_{s \in S_{\mathcal{N}}}$  provides a solution of the following constraints:

$$x_s \geq \text{rew}_{\mathcal{N}}(s, \alpha) + \sum_{s \in S_{\mathcal{N}}} P_{\mathcal{N}}(s, \alpha, t) \cdot x_t \quad \text{for } s \in S \text{ and } \alpha \in \text{Act}_{\mathcal{N}}(s)$$

and  $x_{\text{final}} = 0$ . It is well known that the vector  $(x_s^*)_{s \in S_{\mathcal{N}}}$  where  $x_s^* = E_{\mathcal{N}, s}^{\mathfrak{S}}(\Diamond \text{final})$  provides the unique solution of the above linear constraints that minimizes  $\sum_{s \in S_{\mathcal{N}}} x_s$ . Hence:

$$x_s^{\mathfrak{S}} \geq E_{\mathcal{N}, s}^{\mathfrak{S}}(\Diamond \text{final})$$

for all states  $s$ . On the other hand, for  $\mathfrak{S}(\cdot, r)$  viewed as a scheduler for  $\mathcal{N}$  we have  $x_s^{\mathfrak{S}} = E_{\mathcal{N}, s}^{\mathfrak{S}(\cdot, r)}(\Diamond \text{final})$ . Thus,

$$x_s^{\mathfrak{S}} = E_{\mathcal{N}, s}^{\mathfrak{S}}(\Diamond \text{final})$$

for all states  $s$ . But then for each function  $\mathfrak{P} : S \setminus \{\text{goal}, \text{fail}\} \rightarrow \text{Act}$  with  $\mathfrak{P}(t) \in \text{Act}(t)$  for all  $t$  (viewed as a scheduler for  $\mathcal{N}$ ) we have:

$$x_s^{\mathfrak{S}} = E_{\mathcal{N}, s}^{\mathfrak{P}}(\Diamond \text{final})$$

This yields:

$$\theta_{s, r} - (\vartheta^* - r) \cdot y_{s, r} \geq \theta_{s, r, \mathfrak{P}} - (\vartheta^* - r) \cdot y_{s, r, \mathfrak{P}}$$

for all states  $s \in S \setminus \{\text{goal}, \text{fail}\}$ , all actions  $\alpha \in \text{Act}(s)$  and all functions  $\mathfrak{P}$ . Hence, (\*) holds for any value  $\vartheta^*$  with  $\vartheta \leq \vartheta^* \leq \vartheta_r$ .

It remains to show that if  $\vartheta^* > \vartheta_r$  then (\*) does not holds for at least one pair  $(s, \mathfrak{P})$ . This, however, is obvious as we can pick a pair  $(s, \alpha)$  such that  $y_{s,r} > y_{s,r,\alpha}$  and

$$\vartheta_r = r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}}$$

If  $\vartheta^* > \vartheta_r$  then

$$\vartheta^* > r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}}$$

But then

$$\theta_{s,r} - (\vartheta^* - r) \cdot y_{s,r} < \theta_{s,r,\mathfrak{P}} - (\vartheta^* - r) \cdot y_{s,r,\mathfrak{P}}$$

This completes the proof of the second statement in Lemma G.6. To prove the last statement of Lemma G.6, we suppose

$$\vartheta < \vartheta^* \leq \min \{ \vartheta_R : r \leq R \leq \wp \}$$

(\*) yields that for each level  $R \in \{r, \dots, \wp\}$ , the unique solution of the linear program in Figure 3 is the vector  $(x_s^R)_{s \in S}$  where  $x_s^R = \theta_{s,R} - (\vartheta^* - r)y_{s,R}$ . But then the calls of the threshold algorithm for  $\vartheta$  and  $\vartheta^*$  deal level-wise with the same MDP  $\mathcal{M}_R^* \stackrel{\text{def}}{=} \mathcal{M}_{R,\vartheta}^* = \mathcal{M}_{R,\vartheta^*}^*$  to derive  $y_{s,R} = y_{s,R}^*$  as the maximal probability to reach *goal* from  $s$  in  $\mathcal{M}_R^*$  and  $\theta_{s,R} = \theta_{s,R}^*$  as the expected total reward under each scheduler for  $\mathcal{M}_R^*$ . ■

*Remark G.7 (MDP without zero-reward cycles).* For the special case of an MDP without zero-reward cycles, the presented algorithm for the threshold problem can be simplified as follows. The initialization phase remains unchanged, but in the treatment of the level  $r = \wp - 1, \wp - 2, \dots, 1, 0$ , the solution of the linear program in Figure 3 can be computed directly without linear programming techniques. For this, we consider an enumeration  $s_1, s_2, \dots, s_N$  of the states in  $S \setminus \{\text{goal}, \text{fail}\}$  such that  $P(s_i, \alpha, s_j) > 0$  and  $\text{rew}(s_i, \alpha) = 0$  implies  $i > j$ . Then, for  $i = 1, 2, \dots, N$ , and each action  $\alpha \in \text{Act}(s_i)$  we put

$$y_{s_i,r,\alpha} = \sum_{t \in S} P(s_i, \alpha, t) \cdot y_{t,R}$$

$$\theta_{s_i,r,\alpha} = \text{rew}(s_i, \alpha) \cdot y_{s_i,r,\alpha} + \sum_{t \in S} P(s_i, \alpha, t) \cdot \theta_{t,R}$$

where  $R = \min \{ \wp, r + \text{rew}(s_i, \alpha) \}$ ,  $y_{\text{goal},R} = 1$  and  $y_{\text{fail},R} = \theta_{\text{goal},R} = \theta_{\text{fail},R} = 0$ . Note that if  $\text{rew}(s_i, \alpha) = 0$  then  $R = r$  and  $P(s_i, \alpha, t) > 0$  implies  $t = s_j$  for some  $j < i$ . Hence, the relevant values  $y_{t,R}$  and  $\theta_{t,R}$  have been computed before. Let

$$\Delta_{s_i,r} = \max_{\alpha \in \text{Act}(s_i)} \Delta_{s_i,r,\alpha} \quad \text{where} \quad \Delta_{s_i,r,\alpha} = \theta_{s_i,r,\alpha} - (\vartheta - r) \cdot y_{s_i,r,\alpha}$$

We then pick an action  $\alpha \in \text{Act}(s_i)$  where  $\Delta_{s_i,r,\alpha} = \Delta(s_i, r)$  and  $y_{s_i,r,\alpha} \geq y_{s_i,r,\beta}$  for each action  $\beta \in \text{Act}(s_i)$  with  $\Delta_{s_i,r,\beta} = \Delta(s_i, r)$ . We then define  $(\text{action}(s_i, r), y_{s_i,r}, \theta_{s_i,r})$  as the triple  $(\alpha, y_{s_i,r,\alpha}, \theta_{s_i,r,\alpha})$ . Then, the values  $x_s^* =$

$\Delta_{s,r,\alpha}$  obtained after treating all states at level  $r$  constitute the unique solution of the linear program in Figure 3. Hence, the threshold problem in MDPs without zero-reward cycles is solvable in  $\mathcal{O}(|S| \cdot |Act| \cdot \wp)$  steps. ■

Analogous techniques are applicable to establish a PSPACE upper bound for the threshold problem in acyclic MDPs. This will be shown in Lemma I.4.

## G.2 Complexity of the threshold algorithm

The time complexity of the threshold algorithm is dominated by (i) the computation of the saturation point  $\wp$  as described in Section F and (ii) the linear programs to compute feasible actions for the levels  $r = \wp - 1, \wp - 2, \dots, 1, 0$ . The time complexity of step (i) is pseudo-polynomial in the size of  $\mathcal{M}$  as outlined in Section F. To prove the pseudo-polynomial time bound as stated in Theorem 2, we show that the time complexity of step (ii) is pseudo-polynomial in the size of  $\mathcal{M}$  and polynomial in the logarithmic length of the threshold value  $\vartheta$ . As linear programs are solvable in time polynomial in the number of variables and the total logarithmic lengths of the coefficients in the linear constraints, it suffices to establish a pseudo-polynomial bound for the logarithmic lengths of the probability values  $y_{s,r}$  and the partial expectations  $\theta_{s,r}$  that are computed in the threshold algorithm and used in the linear program in Figure 3 in Appendix G.

Given a non-zero-rational number  $x$ , we refer to the unique coprime integers  $n, d$  with  $x = n/d$  and  $d > 0$  as *the numerator* ( $n$ ) and *the denominator* ( $d$ ) of  $x$ . For  $x = 0$  we say *the denominator* is 1 and *the numerator* is 0.

**Lemma G.8.** *Let  $A = (a_{i,j})_{i,j=1,\dots,m}$  be a non-singular  $m \times m$ -matrix with integer values  $a_{i,j}$  whose logarithmic length is bounded by  $K$ . Let  $b = (b_i)_{i=1,\dots,m}$  be an integer vector where the logarithmic length of the values  $b_i$  is bounded by  $L$ . Let  $x = (x_j)_{j=1,\dots,n}$  be the unique solution of the linear equation system  $Ax = b$ . Then, the  $x_j$ 's are rational numbers and the logarithmic length of the least common multiple  $\text{lcm}$  of the denominators of the values  $x_1, \dots, x_m$  is bounded by  $m \log m + Km$ . Moreover, for all  $j \in \{1, \dots, m\}$ , the logarithmic length of  $x_j \cdot \text{lcm} \in \mathbb{Z}$  is at most  $m \log m + K(m-1) + L$ .*

*Proof.* Let  $A_i$  denote the matrix resulting from  $A$  by replacing the  $i$ -th column with the vector  $b$ . By Cramer's rule we have:

$$x_i = \frac{\det(A_i)}{\det(A)}$$

where  $\det(B)$  denotes the determinant of matrix  $B$ . By the Leibniz formula for determinants we have:

$$\det(A) = \sum_{\sigma \in \text{Perm}_m} \text{sign}(\sigma) \cdot a_{1,\sigma(1)} \cdot a_{2,\sigma(2)} \cdot \dots \cdot a_{m,\sigma(m)}$$

where  $\text{Perm}_m$  denotes the set of permutations of the values  $1, \dots, m$  and  $\text{sign}(\sigma)$  the sign of  $\sigma$ . As  $|a_{i,j}| \leq 2^K - 1$  we get:

$$|\det(A)| < m! \cdot 2^{Km} \leq 2^{m \log m + Km}$$

where we use the fact that  $m! \leq m^m = 2^{m \log m}$ . Likewise, we get for the absolute value of the determinant of  $A_i$ :

$$|\det(A_i)| < m! \cdot 2^{K(m-1)} \cdot 2^L \leq 2^{m \log m + K(m-1) + L}$$

This yields the claim.  $\blacksquare$

**Lemma G.9.** *Notations as in Lemma G.8, except that the values  $a_{i,j}$  and  $b_i$  are rational and the logarithmic lengths of the numerators (resp. denominators) of the values  $a_{i,j}$  are bounded by  $K_n$  (resp.  $K_d$ ), while the logarithmic lengths of the numerators (resp. denominators) of the values  $b_i$  are bounded by  $L_n$  (resp.  $L_d$ ). Then, the values  $x_i$  are rational numbers and the logarithmic length of the least common multiple  $\text{lcm}$  of the denominators is bounded by  $m(\log m + K_n + mK_d + L_d)$ . The logarithmic length of  $x_j \cdot \text{lcm}$  is at most  $m(\log m + K_n + mK_d + L_d) + L_n + mK_d + L_d$ .*

*Proof.* We multiply the  $i$ -th row of  $A$  and  $b_i$  with the least common multiple  $\text{lcm}_i$  of the denominators of the values  $a_{i,j}$  and the denominator of  $b_i$ . Let  $A'x = b'$  be the resulting equation system. Obviously,  $\text{lcm}_i < 2^{mK_d + L_d}$ . Thus, the values  $a'_{i,j}$  of the matrix  $A'$  are integers whose logarithmic length is bounded by  $K = K_n + mK_d + L_d$ . Likewise,  $b = (b'_i)_{i=1,\dots,m}$  is an integer vector and the logarithmic length of the values  $b'_i$  is bounded by  $L = L_n + mK_d + L_d$ .

The unique solution  $x$  of  $Ax = b$  is also the unique solution of  $A'x = b'$ . Thus, the claim follows by Lemma G.8.  $\blacksquare$

We now consider the probability values  $y_{s,r}$  computed in the threshold algorithm and show that their logarithmic lengths are polynomially bounded in  $\wp$  and the size of  $\mathcal{M}$ . Analogous arguments can be provided for the partial expectations  $\theta_{s,r}$ . Let  $y_{s,r} = y'_{s,r}/d_r$  where  $d_r$  is the least common multiple of the denominators of  $y_{s,R}$  where  $s$  ranges over all states in  $S$  and  $R$  over the levels in  $\{r, r+1, \dots, \wp\}$ . Hence, the values  $y'_{s,r}$  are non-negative integers and  $d_r$  is a multiple of  $d_R$  for all  $r < R \leq \wp$ . Let  $k_r$  denote the maximal logarithmic length of the values  $y'_{s,r}$ . The goal is show that  $k_r$  and the logarithmic lengths of the values  $d_r$  are polynomially bounded in  $\wp$  and  $\text{size}(\mathcal{M})$  for all  $r \in \{0, 1, \dots, \wp\}$ .

For  $r = \wp$ , the values  $y_{s,\wp} = \Pr_{\mathcal{M},s}^{\mathfrak{M}}(\Diamond \text{goal})$  can be characterized as the unique solution of a linear equation system  $Ax = b$  where  $A$  is a  $m \times m$ -matrix of the form  $I - P'$  with  $m \leq |S|$ . Here  $I$  is the identity matrix and  $P'$  arises from the transition probability matrix of the Markov chain induced by  $\mathfrak{M}$  by deleting certain columns and rows. Using Lemma G.9 we get that the logarithmic length of the values  $y'_{s,\wp}$  is in  $\mathcal{O}(\text{size}(\mathcal{M})^3)$ . Thus,  $k_\wp$  is in  $\mathcal{O}(\text{size}(\mathcal{M})^3)$ .

Let now  $r < \wp$  and let  $\mathfrak{P}_r^*$  denote the memoryless deterministic scheduler chosen by the threshold algorithm for level  $r$ . For  $t \in S \setminus \{\text{goal}, \text{fail}\}$ , let  $\alpha_t = \mathfrak{P}_r^*(t)$  and  $R_t = \min\{\wp, r + \text{rew}(t, \alpha_t)\}$ . Let  $T_r$  denote the set of states  $t \in S$  such that either  $t \in \{\text{goal}, \text{fail}\}$  or  $\text{rew}(t, \alpha_t) > 0$ . For the states  $t \in T_r \setminus \{\text{goal}, \text{fail}\}$ :

$$y_{t,r} = \sum_{u \in S} P(t, \alpha_t, u) \cdot y_{u,R_t}$$

while for the states  $s \in S \setminus T$ :

$$y_{s,r} = \sum_{t \in T_r} \Pr_{\mathcal{M},s}^{\mathfrak{P}_r^*}(-T_r \cup t) \cdot y_{t,r}$$

As a consequence of Lemma G.9 we obtain a polynomial bound for the logarithmic lengths of the reachability probabilities  $\Pr_{\mathcal{M},s}^{\mathfrak{P}_r^*}(-T_r \cup t)$ . More precisely:

**Corollary G.10.** *There exists a polynomial  $g$  of degree 3 such that for all triples  $(\mathfrak{S}, T, t)$  where  $\mathfrak{S}$  is a memoryless deterministic scheduler for  $\mathcal{M}$ ,  $T \subseteq S$  and  $t \in T$ , there exists a positive integer  $d$  and non-negative integers  $n_s$  for  $s \in S$  with  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(-T \cup t) = n_s/d$  and such that the logarithmic lengths of  $n_s$  and  $d$  are bounded by  $g(\text{size}(\mathcal{M}))$ .*

*Proof.* The claim follows by Lemma G.9 as the non-zero values  $\Pr_{\mathcal{M},s}^{\mathfrak{S}}(-T \cup t)$  can be obtained as the unique solution of a linear equation system of the form  $Ax = b$  where  $A = I - P'$  where  $P'$  arises from the transition probability matrix  $(P(s, \mathfrak{S}(s), t))_{s,t \in S}$  by removing certain columns and rows. Note that  $\text{size}(\mathcal{M})$  is an upper bound for the values  $m, K_d, K_n, L_d, L_n$  in Lemma G.9, no matter how  $\mathfrak{S}, T$  and  $t$  are chosen. ■

In what follows, let  $A$  denote the least common multiple of the denominators  $P_d(s, \alpha, t)$  of the transition probabilities  $P(s, \alpha, t)$  when ranging over all triples  $(s, \alpha, t) \in S \times \text{Act} \times S$  where  $P(s, \alpha, t) > 0$ . Obviously,  $A$  is bounded by the product of the values  $P_d(s, \alpha, t)$ . Hence, the logarithmic length of  $A$  is bounded by  $\text{size}(\mathcal{M})$ . Let  $B_r$  denote the least common multiple of the denominators of the reachability probabilities  $\Pr_{\mathcal{M},s}^{\mathfrak{P}_r^*}(-T_r \cup t)$  when ranging over all triples  $(s, t, r)$  with  $s \in S \setminus T_r$ . The logarithmic length of  $B_r$  is bounded by  $|S|^2 \cdot g(\text{size}(\mathcal{M}))$  where  $g$  is as in Corollary G.10.

For the states  $t \in T_r$  we have:

$$y_{t,r} = \sum_{u \in S} P(t, \alpha_t, u) \cdot \frac{y'_{u,R_t}}{d_{R_t}} = \frac{Y_{t,r}}{A \cdot d_{r+1}}$$

where

$$Y_{t,r} = \sum_{u \in S} A \cdot P(t, \alpha_t, u) \cdot y'_{u,R_t} \cdot \frac{d_{r+1}}{d_{R_t}}$$

Note that  $d_{r+1}/d_{R_t}$  is an integer whose logarithmic length is bounded by the logarithmic length of  $d_{r+1}$ . Obviously, the values  $A \cdot P(t, \alpha_t, u)$  are integers with the logarithmic lengths at most  $2 \cdot \text{size}(\mathcal{M})$ . Hence,  $Y_{t,r} \in \mathbb{N}$ . As  $R_t \geq r+1$ , the logarithmic length of the values  $y'_{u,R_t}$  is bounded by  $k_{r+1}$ . Hence:

$$Y_{t,r} \leq |S| \cdot 2^{2\text{size}(\mathcal{M})} \cdot 2^{k_{r+1}} \cdot d_{r+1}$$

For the states  $s \in S \setminus T_r$  we have:

$$y_{s,r} = \frac{Y_{s,r}}{A \cdot B_r \cdot d_{r+1}} \quad \text{where} \quad Y_{s,r} = \sum_{t \in T_r} B_r \cdot \Pr_{\mathcal{M},s}^{\mathfrak{P}_r^*}(-T_r \cup t) \cdot Y_{t,r}$$



The values  $B_r \cdot \Pr_{\mathcal{M},s}^{\mathfrak{P}_r^*}(\neg T_r \cup t)$  are integers with

$$\log(B_r \cdot \Pr_{\mathcal{M},s}^{\mathfrak{P}_r^*}(\neg T_r \cup t)) \leq |S|^2 \cdot g(\text{size}(\mathcal{M})) + g(\text{size}(\mathcal{M}))$$

Here, we use Corollary G.10. We get  $Y_{s,r} \in \mathbb{N}$  and  $d_r \leq A \cdot B_r \cdot d_{r+1}$ . Thus, there exists a polynomial  $f$  such that:

$$\log(d_r) \leq f(\text{size}(\mathcal{M})) + \log(d_{r+1})$$

and  $d_\varphi \leq f(\text{size}(\mathcal{M}))$ . Hence, the logarithmic lengths of the  $d_r$ 's are polynomially bounded in  $\varphi$  and the size of  $\mathcal{M}$  as we have:

$$\log(d_r) \leq (\varphi - r + 1) \cdot f(\text{size}(\mathcal{M}))$$

The above shows that  $y'_{s,r} \leq Y_{s,r}$  for all states  $s \in S$  and there exists a bivariate polynomial  $h$  such that

$$k_r \leq k_{r+1} + h(\varphi, \text{size}(\mathcal{M}))$$

Recall that  $k_r$  denotes the maximal logarithmic length of the values  $y'_{s,r}$ . We may assume w.l.o.g. that  $k_\varphi \leq h(\varphi, \text{size}(\mathcal{M}))$ . Then:

$$k_r \leq (\varphi - r + 1) \cdot h(\varphi, \text{size}(\mathcal{M}))$$

Hence,  $k_r$  and therefore the logarithmic lengths of the values  $y_{s,r}$  are polynomially bounded in  $\varphi$  and  $\text{size}(\mathcal{M})$ . Analogously, we obtain that the logarithmic lengths of the values  $\theta_{s,r}$  are polynomially bounded in  $\varphi$  and  $\text{size}(\mathcal{M})$ .

We conclude that the representation of the linear programs (in particular the bit-presentation of the coefficients) used in the threshold algorithm (Figure 3 in Appendix G) are polynomially bounded in  $\varphi$ ,  $\text{size}(\mathcal{M})$  and the logarithmic length of the threshold value  $\vartheta$ . As the logarithmic length of the saturation point  $\varphi$  is polynomial in  $\text{size}(\mathcal{M})$  (see Appendix F), the time complexity of the threshold algorithm (Section 4 and Appendix G) is pseudo-polynomial.

## H Computing an optimal scheduler and the maximal conditional expectation

We now address the task to compute  $\mathbb{CE}^{\max}$ . As before, we suppose that (A1), (A2) holds and that  $\mathcal{M}$  has no critical schedulers. By the results of the previous sections, we can formulate a simple algorithm that successively calls the threshold algorithm (see Section G.1) for computing the maximal conditional expectation and an optimal scheduler. The preprocessing is the same as for the threshold algorithm, i.e., the algorithm computes the saturation point  $\varphi$  and the deterministic memoryless scheduler  $\mathfrak{M}$  that is known to provide optimal decisions for all paths  $\pi$  with  $\text{rew}(\pi) \geq \varphi$ . Let  $\text{ThresAlgo}[\vartheta]$  denote the scheduler that is generated by calling the threshold algorithm for the threshold value  $\vartheta$ .

$\mathfrak{S} := \mathfrak{M}$  where  $\mathfrak{M}$  is as in Lemma E.14;  
 REPEAT  
 $\vartheta := \mathbb{CE}^{\mathfrak{S}}; \quad \mathfrak{S} := \text{ThresAlgo}[\vartheta];$   
 UNTIL  $\vartheta = \mathbb{CE}^{\mathfrak{S}};$   
 return  $\vartheta$  as the maximal conditional expectation and an optimal scheduler  $\mathfrak{S}$

Let  $\mathfrak{S}_i$  and  $\vartheta_i$  denote the scheduler resp. threshold value at the end of the  $i$ -th iteration of the repeat-loop. Then,  $\mathfrak{S}_i = \text{ThresAlgo}[\vartheta_{i-1}]$  and  $\vartheta_i = \mathbb{CE}^{\mathfrak{S}_i}$  where  $\vartheta_0 = \mathbb{CE}^{\mathfrak{M}}$ . Hence, all calls of the threshold algorithm are successful in the sense that  $\mathbb{CE}^{\mathfrak{S}_i} \geq \vartheta_{i-1}$ .

Using Corollary G.5 we get the following. The above algorithm generates a strictly increasing sequence of threshold values that are the conditional expectations of some deterministic reward-based scheduler that is memoryless from  $\wp$  on. As the set of the latter is finite and bounded by  $K = |\text{Act}|^{\wp|S|}$  there is some  $k \leq K$  such that

$$\vartheta_0 < \vartheta_1 < \vartheta_2 < \dots < \vartheta_{k-1} = \vartheta_k = \mathbb{CE}^{\max}$$

Hence, the algorithm terminates after at most  $K$  iterations and correctly returns  $\mathbb{CE}^{\max}$  and an optimal scheduler. Together with the pseudo-polynomial time complexity of the threshold algorithm (see Section G.2), this yields a double-exponentially time bounded algorithm for the computation of  $\mathbb{CE}^{\max}$ .

In the sequel, we present a more efficient algorithm for computing  $\mathbb{CE}^{\max}$  that runs in single exponential time. The idea is an iterative scheduler-improvement approach that relies on Lemma G.1 and is interleaved with calls of the threshold algorithm to maintain and successively improve a left-closed and right-open interval  $I = [A, B[$  with  $\mathbb{CE}^{\max} \in I$  together with a scheduler  $\mathfrak{S}$  such that  $\mathbb{CE}^{\mathfrak{S}} \in I$ . Similar to the threshold algorithm, the proposed algorithm for computing  $\mathbb{CE}^{\max}$  operates level-wise and freezes optimal decisions for levels  $r = \wp, \wp-1, \wp-2, \dots, 1, 0$ . More precisely, when level  $r$  has been treated then the current scheduler  $\mathfrak{S}$  is strongly optimal for all level  $\geq r$  in the sense of Definition H.1.

Thanks to Corollary G.5, it is possible that an optimal scheduler is found when treating some level  $r > 0$ , in which case the explicit treatment of levels  $r-1, r-2, \dots, 0$  is skipped. (However, the levels  $< r$  have been treated implicitly in the calls of the threshold algorithms.)

**Definition H.1 (Strong  $r$ -optimality,  $r$ -optimality).** Let  $\mathfrak{S}$  be a reward-based scheduler. Scheduler  $\mathfrak{S}$  is said to be strongly optimal from level  $r$  on, briefly called strongly  $r$ -optimal, if for all states  $s \in S$ , all schedulers  $\mathfrak{T}$  and all  $R \in \mathbb{N}$ ,  $R \geq r$  we have:

$$E_s^{\mathfrak{S} \uparrow R} - (\mathbb{CE}^{\max} - R) \cdot \Pr_s^{\mathfrak{S} \uparrow R}(\Diamond \text{goal}) \geq E_s^{\mathfrak{T}} - (\mathbb{CE}^{\max} - R) \cdot \Pr_s^{\mathfrak{T}}(\Diamond \text{goal})$$

Recall that  $\mathfrak{S} \uparrow R$  denotes the residual scheduler given by  $(\mathfrak{S} \uparrow R)(s, i) = \mathfrak{S}(s, R+i)$ . Hence,  $(\mathfrak{S} \uparrow R) \uparrow i = \mathfrak{S} \uparrow (R+i)$ .

The notion of “optimal from level  $r$  on” is used as before. That is, scheduler  $\mathfrak{S}$  is called optimal from level  $r$  on, briefly called  $r$ -optimal, if  $\mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\max}$

implies  $\mathbb{CE}^{\mathfrak{U}} = \mathbb{CE}^{\max}$  where  $\mathfrak{U}(\pi) = \mathfrak{T}(\pi)$  for each finite path  $\pi$  with  $\text{rew}(\pi) < r$  and  $\mathfrak{U}(\pi) = \mathfrak{S}(\pi)$  for each finite path  $\pi$  with  $\text{rew}(\pi) \geq r$ .

The notion “strongly optimal” will be used instead of strongly 0-optimal. Likewise, a scheduler is briefly called optimal if it is 0-optimal. ■

Clearly,  $\mathfrak{S}$  is strongly  $r$ -optimal if and only if the following conditions hold for all states  $s \in S$ , all schedulers  $\mathfrak{T}$  and all  $R \in \mathbb{N}$ ,  $R \geq r$ :

$$\begin{aligned} R + \frac{E_s^{\mathfrak{S} \uparrow R} - E_s^{\mathfrak{T}}}{\Pr_s^{\mathfrak{S} \uparrow R}(\Diamond \text{goal}) - \Pr_s^{\mathfrak{T}}(\Diamond \text{goal})} &\geq \mathbb{CE}^{\max} && \text{if } \Pr_s^{\mathfrak{S} \uparrow R}(\Diamond \text{goal}) > \Pr_s^{\mathfrak{T}}(\Diamond \text{goal}) \\ R + \frac{E_s^{\mathfrak{S} \uparrow R} - E_s^{\mathfrak{T}}}{\Pr_s^{\mathfrak{S} \uparrow R}(\Diamond \text{goal}) - \Pr_s^{\mathfrak{T}}(\Diamond \text{goal})} &\leq \mathbb{CE}^{\max} && \text{if } \Pr_s^{\mathfrak{S} \uparrow R}(\Diamond \text{goal}) < \Pr_s^{\mathfrak{T}}(\Diamond \text{goal}) \\ E_s^{\mathfrak{S} \uparrow R} &\geq E_s^{\mathfrak{T}} && \text{if } \Pr_s^{\mathfrak{S} \uparrow R}(\Diamond \text{goal}) = \Pr_s^{\mathfrak{T}}(\Diamond \text{goal}) \end{aligned}$$

Hence, by Lemma G.1, each strongly  $r$ -optimal scheduler  $\mathfrak{S}$  is  $r$ -optimal. The reverse direction does not hold in general as some pairs  $(s, r)$  might not be reachable under  $\mathfrak{S}$ .

The existence of strongly optimal schedulers is ensured by the following lemma:

**Lemma H.2.** *The scheduler  $\text{ThresAlgo}[\mathbb{CE}^{\max}]$  is strongly optimal. In particular, if  $\mathbb{CE}^{\text{ThresAlgo}[\vartheta]} = \vartheta$  then  $\text{ThresAlgo}[\vartheta]$  is strongly optimal.*

*Proof.* The first part follows by the fact that strong  $r$ -optimality of a reward-based scheduler  $\mathfrak{S}$  is equivalent to the statement that for all states  $s$ ,  $R \in \mathbb{N}$  with  $R \geq r$  and all schedulers  $\mathfrak{T}$ :

$$E_s^{\mathfrak{S} \uparrow R} - (\mathbb{CE}^{\max} - R) \cdot \Pr_s^{\mathfrak{S} \uparrow R}(\Diamond \text{goal}) \geq E_s^{\mathfrak{T}} - (\mathbb{CE}^{\max} - R) \cdot \Pr_s^{\mathfrak{T}}(\Diamond \text{goal})$$

The claim then follows by Lemma G.6 using an induction on  $i = \wp - R$ .

To see the second part, we suppose  $\mathfrak{S} = \text{ThresAlgo}[\vartheta]$  and  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$ . Then,  $\vartheta = \mathbb{CE}^{\max}$  by Corollary G.5. Thus,  $\mathfrak{S}$  is strongly optimal. ■

*Initialization.* The algorithm starts with the scheduler  $\mathfrak{S} = \text{ThresAlgo}[\mathbb{CE}^{\mathfrak{M}}]$  where  $\mathfrak{M}$  is as in Lemma E.14. If  $\mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\mathfrak{M}}$  then the algorithm immediately terminates on the basis of Corollary G.5. Suppose now that  $\mathbb{CE}^{\mathfrak{S}} > \mathbb{CE}^{\mathfrak{M}}$ . The initial interval is  $I = [A, B[$  where  $A = \mathbb{CE}^{\mathfrak{S}}$  and  $B$  is a strict upper bound for  $\mathbb{CE}^{\max}$ , e.g., the upper bound  $\mathbb{CE}^{\text{ub}}$  (plus some small constant) obtained by the preprocessing explained in Section 3 and Appendix C.4.

*Level-wise scheduler improvement.* The algorithm successively determines optimal decisions for the levels  $r = \wp - 1, \wp - 2, \dots, 1, 0$ . It maintains a left-closed and right-open interval  $I = [A, B[$  and a reward-based scheduler  $\mathfrak{S}$  satisfying the invariance  $\mathbb{CE}^{\max} \in I$  and  $\mathbb{CE}^{\mathfrak{S}} \in I$ . The scheduler  $\mathfrak{S}$  has been obtained by the last successful run of the threshold algorithm for the strict bound “ $> \vartheta$ ” applied to some threshold  $\vartheta \leq A$  in the previous interval. Thus,  $\mathfrak{S} = \text{ThresAlgo}[\vartheta]$  and  $\mathbb{CE}^{\mathfrak{S}} > \vartheta$ .

The treatment of level  $r$  consists of a sequence of scheduler improvement steps where at the same time the interval  $I$  is replaced with proper sub-intervals. The scheduler improvements are obtained by calls of the threshold algorithm “does  $\mathbb{CE}^{\max} \geq \vartheta$  hold?” for some appropriate value  $\vartheta \in I$ . If the scheduler  $\mathfrak{S}$  generated by the threshold algorithm enjoys the property  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$  then the algorithm terminates and returns  $\vartheta$  as the maximal condition expectation and  $\mathfrak{S}$  as an optimal scheduler (see Corollary G.5).

Besides the decisions of  $\mathfrak{S}$  (i.e., the actions  $\mathfrak{S}(s, R)$  for all state-reward pairs  $(s, R)$  where  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and  $R \in \{0, 1, \dots, \wp\}$ ) we shall also need the values

$$\begin{aligned} y_{s,r} &= \Pr_{\mathcal{M},s}^{\mathfrak{S}\uparrow r}(\Diamond \text{goal}) = \Pr_{\mathcal{M}^*,s}^{\max}(\Diamond \text{goal}) \\ \theta_{s,r} &= E_{\mathcal{M},s}^{\mathfrak{S}\uparrow r}(\Diamond \text{goal}) = x_s^* + (\vartheta - r) \cdot y_{s,r} \end{aligned}$$

that have been computed in the threshold algorithm (where  $x_s^*$  refers to the unique solution of the linear program in Figure 3 and  $\mathcal{M}^* = \mathcal{M}_{r,\vartheta}^*$  is the MDP defined as in Section G.1). The algorithm also stores the optimal actions and the values  $\theta_{s,R}$  and  $y_{s,R}$  for  $s \in S$  and all levels  $R \in \{r+1, \dots, \wp\}$  that have been treated before. These values can be reused in the calls of the threshold algorithms. That is, the calls of the threshold algorithm that are invoked in the scheduler-improvement steps at level  $r$  can skip levels  $\wp, \wp-1, \dots, r+1$  and only need to process levels  $r, r-1, \dots, 1, 0$ .

For the current level  $r$ , the algorithm also computes for each state  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and each action  $\alpha \in \text{Act}(s)$  the values:

$$\begin{aligned} y_{s,r,\alpha} &= \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,R} \\ \theta_{s,r,\alpha} &= \text{rew}(s, \alpha) \cdot y_{s,r,\alpha} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_{t,R} \end{aligned}$$

where  $R = \min\{\wp, r + \text{rew}(s, \alpha)\}$ . Thus,  $R = r$  if  $\text{rew}(s, \alpha) = 0$ .<sup>12</sup>

*Scheduler-improvement step.* Let  $r$  be the current level,  $I = [A, B[$  the current interval and  $\mathfrak{S}$  the current scheduler with  $\mathbb{CE}^{\max} \in I$ . At the beginning of the scheduler-improvement step we have  $\mathbb{CE}^{\mathfrak{S}} = A$ . Let

$$\begin{aligned} \mathcal{I}_{\mathfrak{S},r} &= \left\{ r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} : s \in S \setminus \{\text{goal}, \text{fail}\}, \alpha \in \text{Act}(s), y_{s,r} > y_{s,r,\alpha} \right\} \\ \mathcal{I}_{\mathfrak{S},r}^{\uparrow} &= \{ d \in \mathcal{I}_{\mathfrak{S},r} : d \geq \mathbb{CE}^{\mathfrak{S}} \} \quad \mathcal{I}_{\mathfrak{S},r}^B = \{ d \in \mathcal{I}_{\mathfrak{S},r} : d < B \} \end{aligned}$$

If  $\mathcal{I}_{\mathfrak{S},r}^B = \emptyset$  then no further scheduler-improvements at level  $r$  are possible, i.e.,  $\mathfrak{S}$  is strongly  $r$ -optimal (see Lemma H.7). In this case:

<sup>12</sup> Note that  $y_{s,r,\alpha} = \Pr_{\mathcal{M},s}^{\mathfrak{S}_{s,r,\alpha}}(\Diamond \text{goal})$  and  $\theta_{s,r,\alpha} = E_{\mathcal{M},s}^{\mathfrak{S}_{s,r,\alpha}}(\Diamond \text{goal})$ , where  $\mathfrak{S}_{s,r,\alpha}$  denotes the unique scheduler that agrees with  $\mathfrak{S}\uparrow r$ , except that it assigns action  $\alpha$  to state  $s$  (viewed as a path of length 0). That is,  $\mathfrak{S}_{s,r,\alpha}(s) = \alpha$  and  $\mathfrak{S}_{s,r,\alpha}(\pi) = \mathfrak{S}\uparrow r(\pi)$  for all finite paths of length at most 1.

- If  $r = 0$  then the algorithm terminates with the answer  $\mathbb{CE}^{\max} = \mathbb{CE}^{\mathfrak{S}}$  and  $\mathfrak{S}$  as an optimal scheduler.
- If  $r > 0$  then the algorithm goes to the next level  $r-1$  and performs the scheduler-improvement step for  $\mathfrak{S}$  at level  $r-1$ .

Suppose now that  $\mathcal{I}_{\mathfrak{S},r}^B$  is nonempty. Let  $\mathcal{K} = \mathcal{I}_{\mathfrak{S},r}^{\uparrow} \cup \{\mathbb{CE}^{\mathfrak{S}}\}$ . The algorithm seeks for the largest value  $\vartheta' \in \mathcal{K} \cap I$  such that  $\mathbb{CE}^{\max} \geq \vartheta'$ . More precisely, it successively calls the threshold algorithm for the threshold value  $\max(\mathcal{K} \cap I)$  and performs the following steps after each call of the threshold algorithm. Let  $\mathfrak{S}' = \text{ThresAlgo}[\vartheta']$  be the scheduler that has been generated by the threshold algorithm  $\vartheta' = \max(\mathcal{K} \cap I)$ .

- Suppose the result of the threshold algorithm is “no”. If  $\Pr_{\mathcal{M},s_{init}}^{\mathfrak{S}'}(\Diamond \text{goal})$  is positive and  $\mathbb{CE}^{\mathfrak{S}'} \leq \mathbb{CE}^{\max} < \vartheta'$ , then:
  - If  $\mathbb{CE}^{\mathfrak{S}'} \leq A$  then the algorithm refines  $I$  by putting  $B := \vartheta'$ .<sup>13</sup>
  - If  $\mathbb{CE}^{\mathfrak{S}'} > A$  then the algorithm refines  $I$  by putting  $A := \mathbb{CE}^{\mathfrak{S}'}$ ,  $B := \vartheta'$  and adds  $\mathbb{CE}^{\mathfrak{S}'}$  to  $\mathcal{K}$  (Note that then  $\mathbb{CE}^{\mathfrak{S}'} \in \mathcal{K} \cap I$ , while  $\mathbb{CE}^{\mathfrak{S}} \in \mathcal{K} \setminus I$ ).
- Suppose now that the result of the threshold algorithm is “yes”, i.e.,  $\mathbb{CE}^{\mathfrak{S}'} \geq \vartheta'$ . The algorithm terminates if  $\mathbb{CE}^{\mathfrak{S}'} = \vartheta'$ , in which case  $\mathfrak{S}'$  is optimal (Corollary G.5). Otherwise, i.e., if  $\mathbb{CE}^{\mathfrak{S}'} > \vartheta'$ , then the algorithm aborts the loop by putting  $\mathcal{K} := \emptyset$ , refines the interval  $I$  by putting  $A := \mathbb{CE}^{\mathfrak{S}'}$ , updates the current scheduler by setting  $\mathfrak{S} := \mathfrak{S}'$  and repeats the scheduler-improvement step.

The scheduler  $\mathfrak{S}'$  of the last case ( $\mathbb{CE}^{\mathfrak{S}'} \geq \vartheta'$ ) will be called the *outcome* of the scheduler-improvement step for  $\mathfrak{S}$ . Lemma H.4 (see below) shows that the scheduler-improvement step indeed terminates and finds some scheduler  $\mathfrak{S}'$  such that:

$$\mathbb{CE}^{\mathfrak{S}} < \mathbb{CE}^{\mathfrak{S}'} \quad \text{or} \quad \mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\mathfrak{S}'} = \mathbb{CE}^{\max}$$

Let  $\mathfrak{S}_1, \mathfrak{S}_2, \mathfrak{S}_3 \dots$  be the sequence of schedulers  $\mathfrak{S}$  where the algorithm executes a scheduler-improvement step for  $\mathfrak{S}$ . With  $\vartheta_1 = \mathbb{CE}^{\mathfrak{M}}$  and  $\mathfrak{S}_1 = \text{ThresAlgo}[\vartheta_1]$ ,  $\mathfrak{S}_{i+1}$  is the outcome of the scheduler-improvement step for  $\mathfrak{S}_i$  for  $i \geq 1$ . Furthermore, let  $\vartheta_{i+1}$  denote the threshold value such that  $\mathfrak{S}_{i+1}$  is generated by calling the threshold algorithm for  $\vartheta_{i+1}$ . By the choice of the threshold values, we have:

$$\mathbb{CE}^{\mathfrak{S}_{i+1}} > \vartheta_{i+1} \geq \mathbb{CE}^{\mathfrak{S}_i} \quad \text{or} \quad \mathbb{CE}^{\max} = \mathbb{CE}^{\mathfrak{S}_{i+1}} = \vartheta_{i+1} \geq \mathbb{CE}^{\mathfrak{S}_i}$$

All schedulers generated by the algorithm have the form  $\text{ThresAlgo}[\vartheta]$  for some threshold value  $\vartheta$ . In particular, they are deterministic and reward-based schedulers with fixed values for the last level  $\wp$ . The total number of such schedulers

<sup>13</sup> Note that the case  $\mathbb{CE}^{\mathfrak{S}'} < \mathbb{CE}^{\mathfrak{S}}$  is possible, although  $\mathfrak{S}'$  has been generated by the threshold algorithm for a threshold value  $\vartheta'$  that is larger than the threshold  $\vartheta$  used for the generation of  $\mathfrak{S}$ .

is bounded by  $md^\wp$  where  $md$  denotes the total number of memoryless deterministic schedulers of  $\mathcal{M}$ . (Thus,  $md \leq |Act|^{|S|}$ .) Hence, there is some  $k \in \mathbb{N}$  with  $1 \leq k \leq md^\wp + 1$  such that:

$$\mathsf{CE}^{\mathfrak{S}_1} < \mathsf{CE}^{\mathfrak{S}_2} < \mathsf{CE}^{\mathfrak{S}_3} < \dots < \mathsf{CE}^{\mathfrak{S}_{k-1}} < \mathsf{CE}^{\mathfrak{S}_k} \leq \mathsf{CE}^{\mathfrak{S}_{k+1}} = \mathsf{CE}^{\max}$$

This argument yields a proof sketch for the termination and soundness. We now provide a more careful analysis of the algorithm with respect to the correctness and the complexity. Indeed, we will show that the total number of scheduler-improvement steps is in  $\mathcal{O}(\wp \cdot md \cdot |S| \cdot |Act|)$ .

**Theorem H.3 (Soundness and complexity).** *The above algorithm correctly computes  $\mathsf{CE}^{\max}$  and a scheduler  $\mathfrak{S}$  with  $\mathsf{CE}^{\mathfrak{S}} = \mathsf{CE}^{\max}$ . Its time complexity is exponential in the size of the MDP.*

The proof of Theorem H.3 is splitted into two parts. Partial correctness will be shown in Proposition H.8, while the statement on the complexity will be shown in Proposition H.12.

**Lemma H.4 (Correctness of the scheduler-improvement step).** *Let  $\mathfrak{S}$  be a scheduler.*

- (a) *If  $\mathfrak{S}'$  is the outcome of the scheduler-improvement step for  $\mathfrak{S}$  then either  $\mathsf{CE}^{\mathfrak{S}} < \mathsf{CE}^{\mathfrak{S}'}$  or  $\mathsf{CE}^{\mathfrak{S}} = \mathsf{CE}^{\mathfrak{S}'} = \mathsf{CE}^{\max}$  and  $\mathfrak{S}'$  is strongly optimal.*
- (b) *The scheduler-improvement step for  $\mathfrak{S}$  terminates after at most  $|S| \cdot |Act|$  calls of the threshold algorithm.*

*Proof.* We first show statement (a). Obviously,  $\mathfrak{S}'$  is strongly optimal if the scheduler-improvement step terminates on the basis of Corollary G.5 (see also Lemma H.2). Suppose now that  $\mathsf{CE}^{\mathfrak{S}'} < \mathsf{CE}^{\max}$ .

It is obvious that the refinements of the interval  $I = [A, B[$  in the scheduler-improvements are safe in the sense that  $\mathsf{CE}^{\max} \in I$  holds at any moment of the execution of the scheduler-improvement step. When the scheduler-improvement step for  $\mathfrak{S}$  is called then we have  $A = \mathsf{CE}^{\mathfrak{S}}$ .

The set  $\mathcal{K}$  is modified during the execution of the scheduler-improvement step for  $\mathfrak{S}$ . However, at any moment the set  $\mathcal{K}$  only contains elements of  $\mathcal{I}_{\mathfrak{S},r}^\uparrow$  and values of the form  $\mathsf{CE}^{\mathfrak{T}}$  for some scheduler  $\mathfrak{T}$  with  $\mathsf{CE}^{\mathfrak{T}} \geq \mathsf{CE}^{\mathfrak{S}}$ . Furthermore, if  $\mathcal{K}$  is nonempty then  $\mathcal{K} \cap I$  contains exactly one value  $\mathsf{CE}^{\mathfrak{T}}$  for some scheduler  $\mathfrak{T}$  with  $\mathsf{CE}^{\mathfrak{T}} \geq \mathsf{CE}^{\mathfrak{S}}$ . Note that initially we have  $\mathsf{CE}^{\mathfrak{S}} \in \mathcal{K} \cap I$  and whenever an element  $\mathsf{CE}^{\mathfrak{T}}$  is added to  $\mathcal{K}$  then  $\mathfrak{T}$  is the best scheduler found so far and the left border  $A$  of  $I$  is updated to  $A = \mathsf{CE}^{\mathfrak{T}}$ . Furthermore,  $d \geq \mathsf{CE}^{\mathfrak{S}}$  for all  $d \in \mathcal{I}_{\mathfrak{S},r}^\uparrow$ . Thus, as long as  $\mathcal{K}$  is nonempty,  $\mathsf{CE}^{\mathfrak{S}} \leq \min |\mathcal{K} \cap I|$  and  $\mathcal{K} \cap I$  contains at least one value  $\vartheta'$  with  $\mathsf{CE}^{\max} \geq \vartheta'$ .

The refinements of  $I$ 's right border  $B$  ensure that each element of  $\mathcal{I}_{\mathfrak{S},r}^\uparrow$  is selected as the maximal element of  $\mathcal{K} \cap I$  at most once.

Hence, eventually some threshold value  $\vartheta' \in \mathcal{K} \cap I$  with  $\mathsf{CE}^{\max} \geq \vartheta'$  will be picked as the maximal element of  $\mathcal{K} \cap I$ . Let  $\mathfrak{S}' = \text{ThresAlgo}[\vartheta']$ . But then  $\mathsf{CE}^{\mathfrak{S}'} \geq \vartheta'$ . Hence,  $\mathfrak{S}'$  is the outcome of the scheduler-improvement step for  $\mathfrak{S}$ . Moreover:

- If  $\mathbb{CE}^{\mathfrak{S}'} = \vartheta'$  then the scheduler-improvement step for  $\mathfrak{S}$  returns  $\mathfrak{S}'$  as a strongly optimal scheduler.
- If  $\mathbb{CE}^{\mathfrak{S}'} > \vartheta'$  then  $\mathbb{CE}^{\mathfrak{S}'} > \vartheta' = \mathbb{CE}^{\mathfrak{T}} \geq \mathbb{CE}^{\mathfrak{S}}$ .

Hence, the outcome of the scheduler-improvement step for  $\mathfrak{S}$  is a scheduler  $\mathfrak{S}'$  with  $\mathbb{CE}^{\mathfrak{S}'} > \mathbb{CE}^{\mathfrak{S}}$  or  $\mathbb{CE}^{\mathfrak{S}'} = \mathbb{CE}^{\max}$ .

We now turn to the proof of statement (b). The number of calls of the threshold algorithm is bounded by  $|\mathcal{K} \cap I|$  for the initial set  $\mathcal{K} = \mathcal{I}_{\mathfrak{S},r}^{\uparrow} \cup \{\mathbb{CE}^{\mathfrak{S}}\}$  and the current interval  $I = [A, B[$  at the beginning of the scheduler-improvement step for  $\mathfrak{S}$ . (Thus,  $A = \mathbb{CE}^{\mathfrak{S}}$ .) Then:

$$|\mathcal{K} \cap I| \leq |\mathcal{K}| \leq |\mathcal{I}_{\mathfrak{S},r}| + 1 \leq |S| \cdot (|Act| - 1) + 1 \leq |S| \cdot |Act|$$

This completes the proof of Lemma H.4. ■

*Remark H.5 (Behavior for strongly optimal schedulers).* From the moment on where the algorithm for computing  $\mathbb{CE}^{\max}$  has generated a strongly optimal scheduler  $\mathfrak{S}$  on level  $r$  there are two possible cases:

- If  $\mathcal{I}_{\mathfrak{S},i}^B = \emptyset$  for  $i = 0, 1, \dots, r$  then the algorithm terminates without any further calls of the threshold algorithm.
- If  $i \in \{0, 1, \dots, r\}$  is the maximal index such that  $\mathcal{I}_{\mathfrak{S},i}^B \neq \emptyset$  then the algorithm first freezes the decisions for levels  $r, r-1, \dots, i+1$  and then generates the final scheduler  $\mathfrak{S}'$  in the scheduler-improvement step for  $\mathfrak{S}$  at level  $i$  by calling the threshold algorithm for the threshold  $\vartheta' = \mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\max}$ . Thus,  $\mathbb{CE}^{\mathfrak{S}'} = \vartheta'$  and the algorithm terminates on the basis of Corollary G.5.

Hence, the scheduler-improvement step for a strongly optimal scheduler terminates after at most  $|S| \cdot |Act|$  calls of the threshold algorithm (see part (b) of Lemma H.4). ■

*Remark H.6 (Behavior for strongly  $r$ -optimal schedulers).* If  $\mathfrak{S}$  is a strongly  $r$ -optimal scheduler and  $\min \mathcal{I}_{\mathfrak{S},r} = \mathbb{CE}^{\max}$  then the outcome of the scheduler-improvement step for  $\mathfrak{S}$  at level  $r$  is the strongly optimal scheduler  $\mathfrak{S}' = \text{ThresAlgo}[\mathbb{CE}^{\max}]$ . However, for  $r > 0$  there might be strongly  $r$ -optimal schedulers  $\mathfrak{S}$  with

$$\min \mathcal{I}_{\mathfrak{S},r} > \mathbb{CE}^{\max}$$

In this case, if  $B \leq \min \mathcal{I}_{\mathfrak{S},r}$  then  $\mathcal{I}_{\mathfrak{S},r}^B = \emptyset$  and the scheduler-improvement algorithm for  $\mathfrak{S}$  at level  $r$  freezes the values  $\mathfrak{S}(\cdot, r)$  directly without any further calls of the threshold algorithm by switching to level  $r-1$ . If  $B > \min \mathcal{I}_{\mathfrak{S},r}$  then  $\mathcal{I}_{\mathfrak{S},r}^B$  is nonempty and the outcome of the scheduler-improvement step for  $\mathfrak{S}$  at level  $r$  improves  $B$  to some  $B'$  with  $\mathbb{CE}^{\max} < B' \leq \min \mathcal{I}_{\mathfrak{S},r}$ . However, it does not modify the values  $y_{s,r}$  and  $\theta_{s,r}$ . (The latter is a consequence of the second part of Lemma G.6.) Thus, if  $\mathfrak{S}$  is strongly  $r$ -optimal and  $B > \min \mathcal{I}_{\mathfrak{S},r}$  then the outcome the scheduler-improvement step for  $\mathfrak{S}$  is a scheduler  $\mathfrak{S}'$  with  $\mathcal{I}_{\mathfrak{S},r} = \mathcal{I}_{\mathfrak{S}',r}$  and  $\mathcal{I}_{\mathfrak{S}',r}^{B'} = \emptyset$ . ■

**Lemma H.7.** *Let  $r \in \{0, 1, \dots, \wp-1\}$  and  $\vartheta$  a rational value such that  $\vartheta \leq \mathbb{CE}^{\max}$  and  $\mathfrak{S} = \text{ThresAlgo}[\vartheta]$  is strongly  $(r+1)$ -optimal. Then (with  $\min \emptyset = \infty$ ):*

$$\min \mathcal{I}_{\mathfrak{S},r} \geq \mathbb{CE}^{\max} \quad \text{iff} \quad \mathfrak{S} \text{ is strongly } r\text{-optimal}$$

*In particular, if  $\mathcal{I}_{\mathfrak{S},r}^B = \emptyset$  at the beginning of the scheduler-improvement step for  $\mathfrak{S}$  at level  $r$  then  $\mathfrak{S}$  is strongly  $r$ -optimal.*

*Proof.* It is obvious that  $\min \mathcal{I}_{\mathfrak{S},r} \geq \mathbb{CE}^{\max}$  for each strongly  $r$ -optimal scheduler  $\mathfrak{S}$ . We now suppose  $\min \mathcal{I}_{\mathfrak{S},r} \geq \mathbb{CE}^{\max}$  and show the strong  $r$ -optimality of  $\mathfrak{S}$ . Let  $y_{s,r} = \Pr_s^{\mathfrak{S} \uparrow r}(\diamond \text{goal})$  and  $\theta_{s,r} = E_s^{\mathfrak{S} \uparrow r}$ . As  $\vartheta \leq \mathbb{CE}^{\max}$  we have  $\vartheta \leq \mathbb{CE}^{\mathfrak{S}}$ . Using Lemma G.6, we get that for all states  $s$  and for all functions  $\mathfrak{P} : S \setminus \{\text{goal}, \text{fail}\} \rightarrow \text{Act}$  with  $\mathfrak{P}(t) \in \text{Act}(t)$  for all  $t$ :

$$\begin{aligned} r + \frac{\theta_{s,r} - \theta_{s,r,\mathfrak{P}}}{y_{s,r} - y_{s,r,\mathfrak{P}}} &\geq \vartheta && \text{if } y_{s,r} > y_{s,r,\mathfrak{P}} \\ r + \frac{\theta_{s,r} - \theta_{s,r,\mathfrak{P}}}{y_{s,r} - y_{s,r,\mathfrak{P}}} &\leq \vartheta && \text{if } y_{s,r} < y_{s,r,\mathfrak{P}} \\ \theta_{s,r} &\geq \theta_{s,r,\mathfrak{P}} && \text{if } y_{s,r} = y_{s,r,\mathfrak{P}} \end{aligned}$$

By assumption:

$$r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} \geq \mathbb{CE}^{\max}$$

for all states  $s$  and actions  $\alpha \in \text{Act}(s)$  with  $y_{s,r} > y_{s,r,\alpha}$ . With the notations of Lemma G.6 we obtain  $\vartheta_r \geq \mathbb{CE}^{\max} > \vartheta$ . Hence, we can rely on the second part of Lemma G.6 with  $\vartheta^* = \mathbb{CE}^{\max}$  to obtain:

$$\theta_{s,r} - (\mathbb{CE}^{\max} - r) \cdot y_{s,r} \geq \theta_{s,r,\mathfrak{P}} - (\mathbb{CE}^{\max} - r) \cdot y_{s,r,\mathfrak{P}}$$

for all states  $s \in S \setminus \{\text{goal}, \text{fail}\}$  and all functions  $\mathfrak{P}$ . Hence,  $\mathfrak{S}$  is strongly  $r$ -optimal.

If  $\mathcal{I}_{\mathfrak{S},r}^B$  is empty then  $\min \mathcal{I}_{\mathfrak{S},r} \geq B > \mathbb{CE}^{\max}$  where we use the invariance  $\mathbb{CE}^{\max} \in I = [A, B]$ .  $\blacksquare$

**Proposition H.8 (Partial correctness).** *If the algorithm returns the value  $\vartheta$  and the scheduler  $\mathfrak{S}$  then  $\vartheta = \mathbb{CE}^{\max}$  and  $\mathfrak{S}$  is a strongly optimal scheduler.*

*Proof.* The statement of Proposition H.8 is clear in case of an early termination where the algorithm returns a scheduler  $\mathfrak{S} = \text{ThresAlgo}[\vartheta]$  with  $\mathbb{CE}^{\mathfrak{S}} = \vartheta$  (see Corollary G.5). Let us now suppose that the algorithm treats all levels and returns  $\mathbb{CE}^{\mathfrak{S}}$  for some scheduler  $\mathfrak{S}$  considered at level 0 satisfying the constraint  $\mathcal{I}_{\mathfrak{S},0}^B = \emptyset$ . We prove by induction on  $i = \wp - r$  that each scheduler considered in a scheduler-improvement step at level  $r$  is strongly  $(r+1)$ -optimal and that the final scheduler is strongly optimal. This yields  $\mathbb{CE}^{\mathfrak{S}} = \mathbb{CE}^{\max}$ .

For the first level  $r = \wp$  (basis of induction) the claim is clear as  $\mathfrak{M}$  is strongly  $\wp$ -optimal by Lemma E.16. Let now  $r < \wp$  and  $\mathfrak{S}$  be the current scheduler when



the algorithm switches from level  $r$  to  $r-1$ , provided  $r > 0$ , resp. when the treatment of level 0 is completed if  $r = 0$ . Then,  $\mathcal{I}_{\mathfrak{S},r}^B = \emptyset$ . The task is to show that  $\mathfrak{S}$  is strongly  $r$ -optimal. We may rely on the induction hypothesis stating that at the beginning of the treatment of level  $r$ , the current scheduler is strongly  $(r+1)$ -optimal. As the decisions at levels  $r+1, \dots, \wp$  remain unchanged when treating level  $r$ , all schedulers where a scheduler-improvement step is executed at level  $r$  are strongly  $(r+1)$ -optimal. Hence, the remaining task is to prove that for each state  $s \in S$  and each function  $\mathfrak{P} : S \setminus \{\text{goal}, \text{fail}\} \rightarrow \text{Act}$  with  $\mathfrak{P}(t) \in \text{Act}(t)$  we have:

$$\begin{aligned} r + \frac{\theta_{s,r} - \theta_{s,r,\mathfrak{P}}}{y_{s,r} - y_{s,r,\mathfrak{P}}} &\geq \text{CE}^{\max} && \text{if } y_{s,r} > y_{s,r,\mathfrak{P}} \\ r + \frac{\theta_{s,r} - \theta_{s,r,\mathfrak{P}}}{y_{s,r} - y_{s,r,\mathfrak{P}}} &\leq \text{CE}^{\max} && \text{if } y_{s,r} < y_{s,r,\mathfrak{P}} \\ \theta_{s,r} &\geq \theta_{s,r,\mathfrak{P}} && \text{if } y_{s,r} = y_{s,r,\mathfrak{P}} \end{aligned}$$

The above statement is a consequence of Lemma H.7. Hence, the algorithm correctly freezes the values for level  $r$  if  $\mathcal{I}_{\mathfrak{S}_i}^B = \emptyset$ . With  $r = 0$  we get that the final scheduler  $\mathfrak{S}$  is strongly optimal and the returned value is the maximal conditional expectation. ■

We now address the termination and the complexity of the proposed algorithm. We start with statements about the scheduler-improvement steps.

**Definition H.9 (Indices for the variables of the algorithm).** In what follows, we will use the enumeration of the schedulers and threshold values as explained above. I.e.,  $\vartheta_1 = \text{CE}^{\mathfrak{M}}$ ,  $\mathfrak{S}_1 = \text{ThresAlgo}[\vartheta_1]$  and  $\mathfrak{S}_{i+1} = \text{ThresAlgo}[\vartheta_{i+1}]$  is the outcome of the scheduler-improvement step for  $\mathfrak{S}_i$ . Moreover,  $r_i$  and  $I_i = [A_i, B_i[$  denote the current value of the level variable  $r$  resp. the interval  $I = [A, B[$  when the scheduler-improvement step for  $\mathfrak{S}_i$  starts. We often use the fact that  $r_1 = \wp - 1$ ,  $r_{i+1} \in \{r_i, r_i - 1\}$ ,  $A_i = \text{CE}^{\mathfrak{S}_i} < B_i$ , and  $B_1 \geq B_2 \geq B_3 \geq \dots \geq \text{CE}^{\max}$ . The latter follows by a careful inspection of the scheduler-improvement step and the soundness of the threshold algorithm. Furthermore, let  $y_{s,r,i}$  and  $\theta_{s,r,i}$  stand for the current values of  $y_{s,r}$  and  $\theta_{s,r}$  when the scheduler-improvement for  $\mathfrak{S}_i$  at level  $r = r_i$  starts. We then have  $y_{s,r,i} = \Pr_s^{\mathfrak{S}_i \uparrow r}(\diamond \text{goal}) = y_{s,r,\mathfrak{S}_i(\cdot,r)}$  and  $\theta_{s,r,i} = E_s^{\mathfrak{S}_i \uparrow r} \theta_{s,r,\mathfrak{S}_i(\cdot,r)}$  where we use the notations of Section G.1. ■

The decisions of  $\mathfrak{S}_i$  and  $\mathfrak{S}_{i+1}$  at level  $r_i$  might be the same if  $\vartheta_{i+1} > \min \mathcal{I}_{\mathfrak{S}_i, r_i}^\uparrow$ . In this case, however, the next scheduler  $\mathfrak{S}_{i+2}$  will differ at level  $r_i$ . Moreover, if  $\vartheta_{i+1} > \min \mathcal{I}_{\mathfrak{S}_i, r_i}^\uparrow$ , then the  $\mathfrak{S}_i$  and  $\mathfrak{S}_{i+1}$  do not agree at level  $r_i$ . This will be shown in the following lemma.

**Lemma H.10.** *Suppose  $r = r_i = r_{i+1}$  and  $\text{CE}^{\mathfrak{S}_i} > \vartheta_i$ . Then:*

- (a)  $y_{s,r,i} \geq y_{s,r,i+1}$  for all states  $s$ .

- (b) If  $\mathcal{I}_{\mathfrak{S}_{i,r}}^\uparrow \cap I_i$  is empty or  $\vartheta_{i+1} > \min \mathcal{I}_{\mathfrak{S}_{i,r}}^\uparrow$  then there is at least one state  $t$  such that  $y_{t,r,i} > y_{t,r,i+1}$ .
- (c) Suppose  $\vartheta_{i+1} < \min \mathcal{I}_{\mathfrak{S}_{i,r}}^\uparrow$  and  $\mathfrak{S}_i(\cdot, r) = \mathfrak{S}_{i+1}(\cdot, r)$ . Then,  $\mathcal{I}_{\mathfrak{S}_{i+1,r}}^{B_{i+1}} = \emptyset$ .
- (d) Suppose  $\vartheta_{i+1} = \min \mathcal{I}_{\mathfrak{S}_{i,r}}^\uparrow$  and  $\mathfrak{S}_i(\cdot, r) = \mathfrak{S}_{i+1}(\cdot, r)$  and  $\mathbb{CE}^{\mathfrak{S}_{i+1}} > \vartheta_{i+1}$ . Then, there is at least one state  $t$  where  $y_{t,r,i} > y_{t,r,i+2}$ .

*Proof.* We first observe that  $\mathcal{I}_{\mathfrak{S}_{i,r}}^{B_i}$  is nonempty as otherwise  $\mathfrak{S}_i$  would be strongly  $r$ -optimal by Lemma H.7. In what follows, we simply write  $y_{s,r}$  and  $\theta_{s,r}$  rather than  $y_{s,r,i}$  and  $\theta_{s,r,i}$ . Similarly,  $y'_{s,r}$  and  $\theta'_{s,r}$  stand for  $y_{s,r,i+1}$  and  $\theta_{s,r,i+1}$ . The notations  $y_{s,r,\alpha}$  and  $\theta_{s,r,\alpha}$  have the same meaning as in the previous sections, i.e.,  $y_{s,r,\alpha} = \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,r}$  and  $\theta_{s,r,\alpha} = \text{rew}(s, \alpha) \cdot y_{s,r} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_{t,r}$  where  $R = \min\{\emptyset, r + \text{rew}(s, \alpha)\}$ . Furthermore, let  $\vartheta = \vartheta_i$  and  $\vartheta' = \vartheta_{i+1}$ .

Part (a) follows immediately from Lemma G.6 as we have:

$$(\vartheta - r) \cdot (y_{s,r} - y'_{s,r}) \leq \theta_{s,r} - \theta'_{s,r} \leq (\vartheta' - r) \cdot (y_{s,r} - y'_{s,r})$$

As  $\vartheta < \vartheta'$  we obtain  $y_{s,r} \geq y'_{s,r}$  for all states  $s$ .

To prove part (b), we suppose  $\vartheta' > \min \mathcal{I}_{\mathfrak{S}_{i,r}}^\uparrow$  or  $\mathcal{I}_{\mathfrak{S}_{i,r}}^\uparrow \cap I = \emptyset$ . As  $\mathcal{I}_{\mathfrak{S}_{i,r}}^{B_i}$  is nonempty there is at least one state  $s$  and action  $\alpha \in \text{Act}(s)$  such that

$$r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} < \vartheta'$$

and  $y_{s,r} > y_{s,r,\alpha}$ . Hence:

$$\theta_{s,r} - (\vartheta' - r) \cdot y_{s,r} < \theta_{s,r,\alpha} - (\vartheta' - r) \cdot y_{s,r,\alpha}$$

As  $\mathfrak{S}' = \text{ThresAlgo}[\vartheta']$  and using Lemma G.6 we get that there is at least one state  $t \in S$  such that

$$(y_{t,r}, \theta_{t,r}) \neq (y'_{t,r}, \theta'_{t,r})$$

There is some rational number  $\vartheta$  with  $\mathfrak{S} = \text{ThresAlgo}[\vartheta]$  and  $\mathbb{CE}^{\mathfrak{S}} > \vartheta$ . By Lemma G.6:

$$(\vartheta - r) \cdot (y_{s,r} - y'_{s,r}) \leq \theta_{s,r} - \theta_{s,r,i+1} \leq (\vartheta' - r) \cdot (y_{s,r} - y'_{s,r})$$

for all states  $s$ . As  $\vartheta' > \vartheta$  we obtain  $y_{s,r} \geq y'_{s,r}$  for all  $s \in S$ . Moreover,  $y_{s,r} = y'_{s,r}$  implies  $\theta_{s,r} = \theta'_{s,r}$ . But then  $y_{t,r} > y'_{t,r}$ .

For statement (c), we suppose  $\vartheta' < \min \mathcal{I}_{\mathfrak{S}_{i,r}}^\uparrow$  and  $\mathfrak{S}(\cdot, r) = \mathfrak{S}'(\cdot, r)$ . Clearly, then  $\mathcal{I}_{\mathfrak{S},r} = \mathcal{I}_{\mathfrak{S}',r}$  and  $\mathcal{I}_{\mathfrak{S},r}^\uparrow = \mathcal{I}_{\mathfrak{S}',r}^\uparrow$ . By Lemma G.6:

$$r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} \geq \vartheta' \quad \text{if } y_{s,r} > y_{s,r,\alpha}$$

$$r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} \leq \vartheta' \quad \text{if } y_{s,r} < y_{s,r,\alpha}$$

$$\theta_{s,r} \geq \theta_{s,r,\alpha} \quad \text{if } y_{s,r} = y_{s,r,\alpha}$$

for all states  $s$  and actions  $\alpha \in Act(s)$ . As  $\vartheta' \geq \mathbb{CE}^{\max}$  we get  $\mathcal{I}_{\mathfrak{S},r} = \mathcal{I}_{\mathfrak{S},r}^{\uparrow}$ .

We now use the fact that the scheduler-improvement step attempts to find the largest value in  $\vartheta'' \in \mathcal{I}_{\mathfrak{S},r}^{\uparrow}$  such that  $\mathbb{CE}^{\max} \geq \vartheta''$  by successively running the threshold algorithm for the values in  $\mathcal{I}_{\mathfrak{S},r}^{\uparrow} = \mathcal{I}_{\mathfrak{S},r}$  that are still contained in the current interval  $I$ . As  $\mathfrak{S}'$  has been generated by the threshold algorithm for the threshold  $\vartheta'$ , which is strictly less than  $\min \mathcal{I}_{\mathfrak{S},r}$ , we conclude:

$$\mathbb{CE}^{\max} < B_{i+1} < \min \mathcal{I}_{\mathfrak{S},r}$$

Hence, for each state-action pair  $(s, \alpha)$  with  $\alpha \in Act(s)$  and  $y_{s,r} > y_{s,r,\alpha}$  we have:

$$r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} \geq \mathbb{CE}^{\max}$$

For state-action pair  $(s, \alpha)$  with  $y_{s,r} < y_{s,r,\alpha}$  we have:

$$r + \frac{\theta_{s,r} - \theta_{s,r,\alpha}}{y_{s,r} - y_{s,r,\alpha}} \leq \vartheta' < \mathbb{CE}^{\max}$$

This yields:

$$\theta_{s,r} - (\mathbb{CE}^{\max} - r) \cdot y_{s,r} \geq \theta_{s,r,\alpha} - (\mathbb{CE}^{\max} - r) \cdot y_{s,r,\alpha}$$

for all states  $s$  and actions  $\alpha \in Act(s)$ . But then  $\mathfrak{S}_i$  and  $\mathfrak{S}_{i+1}$  are strongly  $r$ -optimal. As  $B_{i+1} < \min \mathcal{I}_{\mathfrak{S},r}$  (see above) and  $\mathcal{I}_{\mathfrak{S},r} = \mathcal{I}_{\mathfrak{S}',r}$  we get  $\mathcal{I}_{\mathfrak{S}',r}^{B_{i+1}} = \emptyset$ .

For statement (d), we first observe that  $\mathfrak{S}_i(\cdot, r) = \mathfrak{S}_{i+1}(\cdot, r)$  implies  $\mathcal{I}_{\mathfrak{S}_i,r} = \mathcal{I}_{\mathfrak{S}_{i+1},r}$ . Although  $\mathcal{I}_{\mathfrak{S}_{i+1},r}^{\uparrow}$  can be a proper superset of  $\mathcal{I}_{\mathfrak{S}_i,r}^{\uparrow}$ , a call of the threshold algorithm for  $\vartheta_{i+1} = \min \mathcal{I}_{\mathfrak{S}_{i+1},r}^{\uparrow}$  is only possible if the calls of the threshold algorithm for the values  $d \in \mathcal{I}_{\mathfrak{S}_i,r}^{\uparrow} \setminus \{\vartheta_{i+1}\}$  were not successful or have been dropped as  $d$  was known to be larger than  $\mathbb{CE}^{\max}$ . Thus, we have  $d \geq B_{i+1}$  for all  $d \in \mathcal{I}_{\mathfrak{S}_i,r}^{\uparrow} \setminus \{\vartheta_{i+1}\}$ . Hence:

$$\mathcal{I}_{\mathfrak{S}_{i+1},r}^{\uparrow} \cap [A_{i+1}, B_{i+1}[ = \{\vartheta_{i+1}\}$$

Recall that the interval  $I$  at the beginning of the scheduler-improvement step for  $\mathfrak{S}_{i+1}$  is  $I_{i+1} = [A_{i+1}, B_{i+1}[$  where  $A_{i+1} = \mathbb{CE}^{\mathfrak{S}_{i+1}}$ . We now can rely on statement (b) for  $\mathfrak{S}_{i+1}$  rather than  $\mathfrak{S}_i$  to derive statement (d). ■

Recall that  $md$  denotes the number of memoryless deterministic schedulers in  $\mathcal{M}$ .

**Lemma H.11.** *The algorithm performs at most  $2 \cdot \wp \cdot md$  scheduler-improvement steps.*

*Proof.* It suffices to show that there are at most  $2 \cdot md$  scheduler-improvement steps at each level  $r \in \{0, 1, \dots, \wp-1\}$ .

By part (a) of Lemma H.10 we get that if  $r = r_i$  then:

$$y_{s,r,i} \geq y_{s,r,i+1} \geq y_{s,r,i+2} \geq \dots$$

for each state  $s \in S$ . We define the following relation  $\triangleleft = \triangleleft_r$  on the set of schedulers  $\mathfrak{S}_i$  where  $r_i = r$ .

$$\mathfrak{S}_i \triangleleft \mathfrak{S}_j \quad \text{iff} \quad \text{there exists some state } t \in S \text{ with } y_{t,r,i} > y_{t,r,j}$$

Clearly,  $\triangleleft$  is transitive and irreflexive. Moreover,  $\mathfrak{S}_i \triangleleft \mathfrak{S}_j$  implies  $\mathfrak{S}_i(\cdot, r) \neq \mathfrak{S}_j(\cdot, r)$ . As a consequence of parts (b), (c) and (d) of Lemma H.10 we get that for each  $i$  with  $r = r_i = r_{i+1}$ :

$$\begin{aligned} & \mathfrak{S}_i \triangleleft \mathfrak{S}_{i+1} \\ \text{or } & \mathfrak{S}_i(\cdot, r) = \mathfrak{S}_{i+1}(\cdot, r) \text{ and } r = r_{i+2} \text{ and } \mathfrak{S}_i \triangleleft \mathfrak{S}_{i+2} \\ \text{or } & \mathfrak{S}_i(\cdot, r) = \mathfrak{S}_{i+1}(\cdot, r) \text{ and } r_{i+2} = r-1 \text{ (if } r > 0 \text{) resp. the algorithm} \\ & \text{terminates with } \mathfrak{S}_{i+1} \text{ as a strongly optimal scheduler (if } r = 0 \text{)} \end{aligned}$$

Each of the function  $\mathfrak{S}_i(\cdot, r)$  can be viewed as a memoryless deterministic scheduler for  $\mathcal{M}$ . The above shows that each memoryless deterministic scheduler for  $\mathcal{M}$  appears at most twice in the sequence  $\mathfrak{S}_i(\cdot, r), \mathfrak{S}_{i+1}(\cdot, r), \mathfrak{S}_{i+2}(\cdot, r), \dots$  induced by schedulers  $\mathfrak{S}_j$  where the algorithm performs a scheduler-improvement step at level  $r$ . This yields the claim.  $\blacksquare$

**Proposition H.12 (Complexity).** *The algorithm terminates after at most  $2 \cdot \wp \cdot |S| \cdot |Act|^{|S|+1}$  calls of the threshold algorithm. The time complexity is exponential in the size of the MDP.*

*Proof.* The statement follows by a combination of Lemma H.11, part (b) of Lemma H.4 and the results of Section G.2 and using the fact  $md \leq |Act|^{|S|}$ .  $\blacksquare$

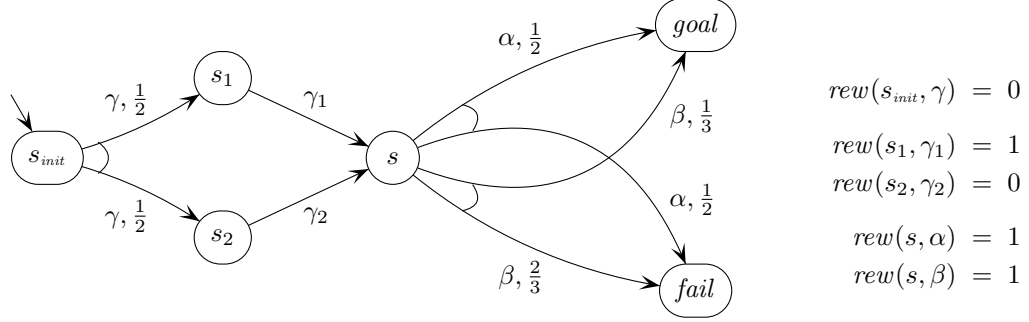
## I PSPACE completeness for acyclic MDPs

We now address the complexity of the four variants of the threshold problem for maximal conditional expectations in MDPs. The pseudo-polynomial algorithms for the threshold problems presented in Section 4 and Appendix G yield an exponential upper bound. The purpose of this section is to show that the threshold problems are PSPACE-complete for acyclic MDPs.

We start with the observation that even for acyclic MDPs history-dependent schedulers can be more powerful to maximize or minimize conditional expected accumulated rewards. Obviously, if  $\mathcal{M}$  is acyclic then  $\mathcal{M}$  enjoys conditions (A1) and (A2) and has no critical scheduler. Thus, the maximal conditional expected accumulated reward for reaching *goal* is finite.

*Example I.1 (History needed in acyclic MDPs).* We regard the acyclic MDP shown in Figure 4. The memoryless schedulers that select always  $\alpha$  resp.  $\beta$  for state  $s$  have the conditional expectation  $3/2$ :

$$\begin{aligned} \text{CE}^{\mathfrak{S}_\alpha} &= \frac{\frac{1}{4} \cdot 2 + \frac{1}{4} \cdot 1}{\frac{1}{4} + \frac{1}{4}} = \frac{\frac{3}{4}}{\frac{1}{2}} = \frac{3}{2} \\ \text{CE}^{\mathfrak{S}_\beta} &= \frac{\frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 1}{\frac{1}{6} + \frac{1}{6}} = \frac{\frac{1}{2}}{\frac{1}{3}} = \frac{3}{2} \end{aligned}$$



**Fig. 4.** MDP  $\mathcal{M}$  for Example I.1

The scheduler  $\mathfrak{T}$  that chooses  $\alpha$  for the path  $\pi_1 = s_{init} \gamma s_1 \gamma_1 s$  and  $\beta$  for the path  $\pi_2 = s_{init} \gamma s_2 \gamma_2 s$  has the conditional expectation  $8/5$  as we have:

$$\mathbb{CE}^{\mathfrak{T}} = \frac{\frac{1}{4} \cdot 2 + \frac{1}{6} \cdot 1}{\frac{1}{4} + \frac{1}{6}} = \frac{\frac{8}{12}}{\frac{5}{12}} = \frac{8}{5}$$

Finally, we regard the scheduler  $\mathfrak{U}$  that chooses  $\beta$  for  $\pi_1$  and  $\alpha$  for  $\pi_2$ . Its conditional expectation is:

$$\mathbb{CE}^{\mathfrak{U}} = \frac{\frac{1}{6} \cdot 2 + \frac{1}{4} \cdot 1}{\frac{1}{6} + \frac{1}{4}} = \frac{\frac{7}{12}}{\frac{5}{12}} = \frac{7}{5}$$

As the maximal conditional expectation is achieved by a deterministic scheduler (see Section D) we get:

$$\frac{7}{5} = \mathbb{CE}^{\min} = \mathbb{CE}^{\mathfrak{U}} < \underbrace{\mathbb{CE}^{\mathfrak{S}_\alpha} = \mathbb{CE}^{\mathfrak{S}_\beta}}_{=\frac{3}{2}} < \mathbb{CE}^{\mathfrak{T}} = \mathbb{CE}^{\max} = \frac{8}{5}$$

Thus, the history-dependent schedulers are superior when the task is to maximize or minimize the conditional expectations.  $\blacksquare$

**Theorem I.2 (PSPACE-completeness for acyclic MDPs).** *All four variants of the threshold problem for maximal conditional expectations in acyclic MDPs are PSPACE-complete.*

As PSPACE is closed under complementation, it suffices to consider the cases where  $\vartheta$  serves as a strict or non-strict lower bound. The proof of Theorem I.2 for lower bounds “ $\geq \vartheta$ ” resp. “ $> \vartheta$ ” is splitted into two parts. The proof for the PSPACE-hardness will be provided in Lemma I.3. Membership to PSPACE will be shown in Lemma I.4.

**Lemma I.3 (PSPACE-hardness for acyclic MDPs).** *The threshold problem for maximal conditional expectations in acyclic MDPs is PSPACE-hard.*

*Proof.* We first address the case where the given threshold value  $\vartheta$  is a non-strict lower bound for the maximal conditional expectations. We provide a polynomial reduction from the problem

given: an acyclic MDP  $\mathcal{N}$  with initial state  $s_0$  and a trap state  $final$  such that  $\Pr_{\mathcal{N},s_0}^{\min}(\Diamond final) = 1$  and a natural number  $R$   
 question: does  $\Pr_{\mathcal{N},s_0}^{\max}(\Diamond^{\geq R} final) \geq \frac{1}{2}$  hold ?

PSPACE-completeness of the above problem has been shown by Haase and Kiefer (Theorem 7 in [25]). In the following, we provide a polynomial reduction that transforms  $\mathcal{N}$  into an acyclic MDP  $\mathcal{M}$  with non-negative rational rewards and threshold  $\vartheta \in \mathbb{Q}$  such that

$$\Pr_{\mathcal{N},s_0}^{\max}(\Diamond^{\geq R} final) \geq \frac{1}{2} \quad \text{iff} \quad \text{CE}_{\mathcal{M},s_{init}}^{\max}(\Diamond goal \mid \Diamond goal) \geq \vartheta$$

In Section J.1 we explain a general approach for transforming MDPs with rational rewards into MDPs with integer rewards of the same asymptotic size. An alternative approach will be sketched at the end of the proof.

In the sequel, let  $S_{\mathcal{N}}$  be the state space of  $\mathcal{N}$ ,  $Act_{\mathcal{N}}$  the action set,  $P_{\mathcal{N}} : S_{\mathcal{N}} \times Act_{\mathcal{N}} \times S_{\mathcal{N}} \rightarrow [0, 1]$  the transition probability function,  $rew_{\mathcal{N}} : S_{\mathcal{N}} \times Act_{\mathcal{N}} \rightarrow \mathbb{N}$  the reward function and  $s_0 \in S_{\mathcal{N}}$  the initial state of  $\mathcal{N}$ . Furthermore, there is a distinguished trap state  $final \in S_{\mathcal{N}}$  (i.e.,  $Act_{\mathcal{N}}(final) = \emptyset$ ) with  $\Pr_{\mathcal{N},s_0}^{\min}(\Diamond final) = 1$ , i.e.,  $\Pr_{\mathcal{N},s_0}^{\mathfrak{T}}(\Diamond final) = 1$  for all schedulers  $\mathfrak{T}$  for  $\mathcal{N}$ . As  $\mathcal{N}$  is acyclic this means that all maximal paths in  $\mathcal{N}$  end in state  $final$ .

The idea of the reduction is to define  $\mathcal{M}$  as the MDP that results from the given MDP  $\mathcal{N}$  by adding a fresh starting state  $s_{init}$ , an auxiliary state  $t$  and two trap states  $goal$  and  $fail$ . In the initial state  $s_{init}$ ,  $\mathcal{M}$  behaves purely probabilistically and moves with probability  $p$  to the new state  $t$  and with probability  $1-p$  to the initial state  $s_0$  of  $\mathcal{N}$ . The reward of the initialization step is 0. From state  $t$ ,  $\mathcal{M}$  moves deterministically to state  $goal$  while earning some reward  $T$ . The probability value  $p$  will be chosen in such a way that the conditional expectation of each scheduler for  $\mathcal{M}$  is between  $T - \frac{1}{4}$  and  $T + \frac{1}{4}$ . As soon as  $\mathcal{N}$  has reached  $s_0$ ,  $\mathcal{M}$  behaves as  $\mathcal{N}$ . For the final state  $final$  of  $\mathcal{N}$ ,  $\mathcal{M}$  offers  $K+1$  actions, called *reject* and *accept*<sub>0</sub>, *accept*<sub>1</sub>, ..., *accept*<sub>K</sub>. The number  $K$  will be chosen in such a way that  $2^{K-1} \leq rew_{\mathcal{N}}(\pi) - R < 2^K$  for all paths  $\pi$  in  $\mathcal{N}$ . With action *reject*,  $\mathcal{M}$  moves with probability 1 and reward 0 to state *fail*. Intuitively, the reject action should be the optimal one for all paths from  $s_{init}$  to *final* with reward less than  $R$ . The actions *accept*<sub>i</sub> for  $0 \leq i < K$  are probabilistic and lead from *final* to *goal* with probability  $\lambda^{K-i}$  and with probability  $1 - \lambda^{K-i}$  to state *fail* for some rational value  $\lambda \in ]0, 1[$ . The reward of *accept*<sub>i</sub> is some value  $X_i$ . When selecting *accept*<sub>K</sub> in state *final*,  $\mathcal{M}$  moves deterministically to state *goal*, while earning reward  $X_K$ . The parameter  $\lambda$  and the reward values  $X_0, \dots, X_K$  will be chosen in such a way that action *accept*<sub>i</sub> is optimal for all paths  $\pi$  from  $s_{init}$  to *final* with  $R - 1 + 2^i \leq rew_{\mathcal{M}}(\pi) < R - 1 + 2^{i+1}$ .

Assuming that appropriate values  $p, \lambda, T, X_0, \dots, X_K$  have been defined, the formal definition of  $\mathcal{M}$  is as follows. The state space of  $\mathcal{M}$  is

$$S_{\mathcal{M}} = S_{\mathcal{N}} \cup \{goal, fail, s_{init}, t\}$$

The action set of  $\mathcal{M}$  is

$$Act_{\mathcal{M}} = Act_{\mathcal{N}} \cup \{\tau, reject, accept_0, \dots, accept_K\}$$

We have  $Act(s_{init}) = Act(t) = \{\tau\}$  and

$$P_{\mathcal{M}}(s_{init}, \tau, t) = p, \quad P_{\mathcal{M}}(s_{init}, \tau, s_0) = 1-p, \quad rew_{\mathcal{M}}(s_{init}, \tau) = 0$$

and

$$P_{\mathcal{M}}(t, \tau, goal) = 1, \quad rew_{\mathcal{M}}(t, \tau) = T$$

For the states  $s \in S_{\mathcal{N}} \setminus \{final\}$  we have  $Act_{\mathcal{M}}(s) = Act_{\mathcal{N}}(s)$  and  $P_{\mathcal{M}}(s, \alpha, t) = P_{\mathcal{N}}(s, \alpha, t)$  and  $rew_{\mathcal{M}}(s, \alpha) = rew_{\mathcal{N}}(s, \alpha)$  for all states  $t \in S_{\mathcal{N}}$  and actions  $\alpha \in Act_{\mathcal{N}}(s)$ . States *goal* and *fail* are trap states in  $\mathcal{M}$ , i.e.,  $Act_{\mathcal{M}}(goal) = Act_{\mathcal{M}}(fail) = \emptyset$ . For the state *final* we have

$$Act_{\mathcal{M}}(final) = \{reject\} \cup \{accept_0, accept_1, \dots, accept_K\}$$

and  $P_{\mathcal{M}}(final, reject, fail) = 1$ ,  $rew_{\mathcal{M}}(final, reject) = 0$  and

$$\begin{aligned} P_{\mathcal{M}}(final, accept_i, goal) &= \lambda^{K-i} \\ P_{\mathcal{M}}(final, accept_i, fail) &= 1 - \lambda^{K-i} \end{aligned} \quad rew_{\mathcal{M}}(final, accept_i) = X_i$$

for  $i = 0, 1, \dots, K$ . In all remaining cases, we have  $P_{\mathcal{M}}(\cdot) = 0$ .

*Some auxiliary notations and choice of  $\lambda$ .* For  $i \in \mathbb{N}$  we define

$$R_i = R - 1 + 2^i$$

Note that  $R_0 = R$  and  $R_{i+1} = R_0 + 2^{i+1} = R_i + 2^i$ . Let

$$E = \sum_{\substack{s \in S \\ s \neq final}} rew_{\mathcal{N}}^{\max}(s)$$

where

$$rew_{\mathcal{N}}^{\max}(s) = \max \{ rew_{\mathcal{N}}(s, \alpha) : \alpha \in Act_{\mathcal{N}}(s) \}$$

Clearly,  $rew_{\mathcal{N}}(\pi) \leq E$  for all paths  $\pi$  in  $\mathcal{N}$ . W.l.o.g.  $R < E$ . Let  $K \in \mathbb{N}$  such that

$$2^{K-1} \leq E - R < 2^K$$

For each state  $s$  in  $\mathcal{N}$ , let  $m_s$  be the least common multiple of the denominators of the probability values  $P_{\mathcal{N}}(s, \alpha, t)$  for some  $\alpha \in Act_{\mathcal{N}}(s)$  and  $\alpha$ -successor  $t$  of  $s$ . Let

$$m = \prod_{\substack{s \in S_{\mathcal{N}} \\ s \neq final}} m_s$$

Then,  $m$  is a natural number and the number of digits of  $m$  in a binary (or decimal) encoding is bounded by the size of  $\mathcal{N}$ . Moreover, as  $\mathcal{N}$  is acyclic, for each maximal path  $\pi$  from  $s_{init}$  to *final*, the probability  $\text{prob}(\pi)$  can be written in

the form  $\ell/m$  for some natural number  $\ell$ . This yields that for any deterministic scheduler  $\mathfrak{S}$  for  $\mathcal{N}$  we have:

$$\Pr_{\mathcal{N}, s_{init}}^{\mathfrak{S}}(\Diamond^{\geq R} final) \in \left\{ \frac{k}{m} : k \in \mathbb{N} \right\} \quad (*)$$

The value  $\lambda$  is defined by:

$$\lambda = 1 - \frac{1}{2Km}$$

By the Bernoulli inequality:

$$\lambda^K = \left(1 - \frac{1}{2Km}\right)^K \geq 1 - \frac{K}{2Km} = 1 - \frac{1}{2m}$$

*Reward parameters*  $X_0, X_1, \dots, X_K$ . The reward values  $X_0, X_1, \dots, X_K$  are defined inductively by:

$$X_i = (1-\lambda) \cdot (X - 2^i + 1) + \lambda X_{i-1} \quad \text{for } i = 1, \dots, K$$

where the choice of  $X_0 = X$  will be explained below. For the following statements up to and including Claim 2, it suffices to deal with any value  $X$  such that  $X \geq 2^K$  and  $X \geq 2Km = 1/(1-\lambda)$ . Note that the constraint  $X \geq 2^K$  yields

$$X_0 = X > X_1 > X_2 > \dots > X_K > 0.$$

Moreover:

$$\begin{aligned} X_i &= \lambda^i X + (1-\lambda) \cdot \sum_{j=0}^i \lambda^j (X - 2^{i-j} + 1) \\ &= \lambda^i X + (1-\lambda) \cdot \sum_{j=0}^i \lambda^j (X + 1) - (1-\lambda) 2^i \cdot \sum_{j=0}^i \left(\frac{\lambda}{2}\right)^j \\ &= \lambda^i X + (1-\lambda)(X + 1) \cdot \frac{1 - \lambda^{i+1}}{1 - \lambda} - (1-\lambda) 2^i \cdot \frac{1 - \left(\frac{\lambda}{2}\right)^{i+1}}{1 - \frac{\lambda}{2}} \\ &= \lambda^i X + (1 - \lambda^{i+1})(X + 1) - \frac{1 - \lambda}{2 - \lambda} \cdot (2^{i+1} - \lambda^{i+1}) \\ &= X + \lambda^i (1 - \lambda) X - \lambda^{i+1} - \frac{1 - \lambda}{2 - \lambda} \cdot (2^{i+1} - \lambda^{i+1}) \\ &= X + \lambda^i (1 - \lambda) X - \frac{1 - \lambda}{2 - \lambda} \cdot 2^{i+1} - \frac{\lambda^{i+1}}{2 - \lambda} \end{aligned}$$

Recall that we require  $X \geq 2Km$ . This implies:

$$X > \frac{1 - \frac{1}{2Km}}{1 + \frac{1}{2Km}} \cdot 2Km = \frac{\lambda}{2 - \lambda} \cdot \frac{1}{1 - \lambda}$$



Hence,  $(1-\lambda)X > \lambda/(2-\lambda)$ . Therefore:

$$\lambda^i(1-\lambda)X > \frac{\lambda^{i+1}}{2-\lambda}$$

This yields

$$X_i > X - \frac{1-\lambda}{2-\lambda} \cdot 2^{i+1}$$

and therefore  $X_i + 2^i > X$  as we have:

$$\begin{aligned} X_i + 2^i &> X - \frac{1-\lambda}{2-\lambda} \cdot 2^{i+1} + 2^i \\ &= X + 2^i \cdot \left( 1 - 2 \cdot \frac{1-\lambda}{2-\lambda} \right) \\ &= X + 2^i \cdot \frac{2-\lambda-2+2\lambda}{2-\lambda} \\ &= X + \frac{2^{i+1}\lambda}{2-\lambda} > X \end{aligned}$$

Furthermore, the inductive definition of the values  $X_i$  yields:

$$\Delta_i \stackrel{\text{def}}{=} \frac{\lambda^{K-i}X_i - \lambda^{K-i+1}X_{i-1}}{\lambda^{K-i} - \lambda^{K-i+1}} = \frac{X_i - \lambda X_{i-1}}{1-\lambda} = X - 2^i + 1$$

*Definition of the parameters of the initialization.* We define

$$T = R + X - \frac{1}{2}$$

and choose a rational number  $p \in ]0, 1[$  such that

$$pT \geq T - \frac{1}{4} = R + X - \frac{3}{4}$$

and

$$\frac{1}{p} \cdot \left( pT + (1-p)(R + 2^K + X) \right) \leq T + \frac{1}{4} = R + X - \frac{1}{4}$$

The latter constraint is equivalent to

$$\frac{1-p}{p} \cdot (R + 2^K + X) \leq \frac{1}{4}$$

which again is equivalent to

$$\frac{1}{p} \cdot (R + 2^K + X) \leq \frac{1}{4} + R + 2^K + X$$

For example, we can deal with

$$p = \max \left\{ \frac{R + 2^K + X}{\frac{1}{4} + R + 2^K + X}, \frac{R + X - \frac{3}{4}}{R + X - \frac{1}{2}} \right\}$$

Note that then indeed  $0 < p < 1$ . Moreover, the choice of  $p$  ensures that:

$$T - \frac{1}{4} \leq \mathbb{CE}^{\mathfrak{S}} \leq T + \frac{1}{4}$$

for each scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  with  $\Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal) > 0$ . In particular:

$$\begin{aligned} \mathbb{CE}^{\max} &\leq T + \frac{1}{4} = R + X - \frac{1}{4} < R + X \\ \mathbb{CE}^{\min} &\geq T - \frac{1}{4} = R + X - \frac{3}{4} > R + X - 1 \end{aligned}$$

*Optimal decisions in the final state.* As before, if  $\mathfrak{S}$  is a scheduler for  $\mathcal{M}$  then we write  $\mathbb{CE}^{\mathfrak{S}}$  for  $\mathbb{CE}_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\Diamond goal | \Diamond goal)$ . We use here the fact that  $rew_{\mathcal{M}}(\pi) \leq E \leq R + 2^K$  for each path from  $s_{init}$  to *final* in  $\mathcal{M}$  and  $X_i \leq X$  for  $i = 0, 1, \dots, K$ . By the choice of the reward values  $X_0, \dots, X_K$ , we obtain that action *accept<sub>i</sub>* is optimal for each finite path  $\pi$  from  $s_{init}$  to *final* with  $R_i \leq rew_{\mathcal{M}}(\pi) < R_{i+1}$ . To see this, we rely on Lemma E.3 and the observation:

$$\underbrace{rew_{\mathcal{M}}(\pi)}_{\geq R_i} + \Delta_i \geq R_i + X - 2^i + 1 = R + X > \mathbb{CE}^{\max}$$

Thus, if  $rew_{\mathcal{M}}(\pi) \geq R_i$  then action *accept<sub>i</sub>* yields a better (larger) conditional expectation than *accept<sub>i-1</sub>*. Likewise, for the paths  $\pi$  from  $s_{init}$  to *final* with  $rew(\pi) < R_i$ , action *accept<sub>i-1</sub>* is better than *accept<sub>i</sub>* as we have:

$$\underbrace{rew_{\mathcal{M}}(\pi)}_{\leq R_i - 1} + \Delta_i \leq R_i - 1 + X - 2^i + 1 = R + X - 1 < \mathbb{CE}^{\min}$$

Action *reject* is the optimal one for exactly the paths  $\pi$  from  $s_{init}$  to *goal* with  $rew(\pi) < R$  as we have:

$$\underbrace{rew_{\mathcal{M}}(\pi)}_{\leq R-1} + \frac{\lambda^{K-i} X_i - 0}{\lambda^{K-i} - 0} \leq R - 1 + X_i \leq R + X - 1 < \mathbb{CE}^{\min}$$

Let *SchedOpt* denote the class of scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  such that for each finite paths  $\pi$  from  $s_{init}$  to *final* we have:

$$\mathfrak{S}(\pi) = \begin{cases} \textit{accept}_i & : \text{ if } R_i \leq rew_{\mathcal{M}}(\pi) < R_{i+1} \\ \textit{reject} & : \text{ if } rew_{\mathcal{M}}(\pi) < R \end{cases}$$

The above shows that for each scheduler  $\mathfrak{U}$  for  $\mathcal{M}$  there is a scheduler  $\mathfrak{S} \in \text{SchedOpt}$  with  $\mathbb{CE}^{\mathfrak{S}} \geq \mathbb{CE}^{\mathfrak{U}}$ .

*Functions  $f$  and  $g$ .* We define functions  $f, g : [0, 1] \rightarrow \mathbb{R}$  as follows:

$$\begin{aligned} f(q) &= \frac{pT + (1-p)q\lambda^K(R+X)}{p + (1-p)q\lambda^K} \\ g(q) &= \frac{pT + (1-p)q(R+2^K+X_K)}{p + (1-p)q} \end{aligned}$$

Both  $f$  and  $g$  are strictly monotonous as their first derivatives  $f'$  and  $g'$  are positive. Note that

$$\text{if } h(q) = \frac{a+bq}{c+dq} \text{ then } h'(q) = \frac{bc-ad}{(c+dq)^2}$$

In the case of  $f$ , we deal with  $a = pT$ ,  $b = (1-p)\lambda^K(R+X)$ ,  $c = p$  and  $d = (1-p)\lambda^K$ . Then:

$$\begin{aligned} bc - ad &= (1-p)\lambda^K(R+X)p - pT(1-p)\lambda^K \\ &= p(1-p)\lambda^K(R+X-T) \\ &= p(1-p)\lambda^K \cdot \frac{1}{2} > 0 \end{aligned}$$

In the case of  $g$  we deal with  $a = pT$ ,  $b = (1-p)(R+2^K+X_K)$ ,  $c = p$  and  $d = 1-p$ . Then:

$$\begin{aligned} bc - ad &= (1-p)(R+2^K+X_K)p - pT(1-p) \\ &= p(1-p)(R+2^K+X_K-T) > 0 \end{aligned}$$

Note that  $2^K+X_K > X$  and  $T = R+X-\frac{1}{2} < R+X$ , which yields  $R+2^K+X_K-T > 0$ .

*Claim 1:* For each scheduler  $\mathfrak{S} \in \text{SchedOpt}$  for  $\mathcal{M}$  with  $q = \Pr_{\mathcal{N},s_0}^{\mathfrak{S}}(\Diamond \text{final})$  we have:

$$f(q) \leq \mathbb{CE}^{\mathfrak{S}} \leq g(q)$$

*Proof of Claim 1:* As before, we write  $\mathbb{CE}^{\mathfrak{S}}$  for  $\mathbb{CE}_{\mathcal{M},s_{init}}^{\mathfrak{S}}(\Diamond \text{goal} | \Diamond \text{goal})$ . Let

$$q_r = \Pr_{\mathcal{M},s_{init}}^{\mathfrak{S}}(\Diamond^{=r} \text{final})$$

and

$$A = \sum_{i=0}^K \sum_{r=R_i}^{R_{i+1}-1} q_r \cdot \lambda^{K-i} \cdot (r + X_i)$$

Then:

$$\mathbb{CE}^{\mathfrak{S}} = \frac{pT + (1-p)A}{p + (1-p)(q_0\lambda^K + q_1\lambda^{K-1} + \dots + q_K)}$$

We first show that  $f(q) \leq \mathbb{CE}^{\mathfrak{S}}$ . Thanks to Lemma E.3, it suffices to show that for each  $i \in \{1, \dots, K-1\}$  and  $r \in \mathbb{N}$  with  $R_i \leq r < R_{i+1}$  we have:

$$\frac{\lambda^{K-i}(r + X_i) - \lambda^K(R + X)}{\lambda^{K-i} - \lambda^K} \geq \mathbb{CE}^{\max}$$

To prove this, we show by induction on  $i$ :

$$\frac{X_i - \lambda^i X}{1 - \lambda^i} \geq X - 2^i + 1$$

The case  $i = 1$  is obvious as

$$\frac{X_1 - \lambda X}{1 - \lambda} = \frac{X_1 - \lambda X_0}{1 - \lambda} = \Delta_1 = X - 1$$

Induction step:

$$\begin{aligned} \frac{X_i - \lambda^i X}{1 - \lambda^i} &= \frac{X_i - \lambda X_{i-1}}{1 - \lambda^i} + \frac{\lambda X_{i-1} - \lambda^i X}{1 - \lambda^i} \\ &= \underbrace{\frac{X_i - \lambda X_{i-1}}{1 - \lambda}}_{=X-2^{i-1}+1} \cdot \frac{1 - \lambda}{1 - \lambda^i} + \underbrace{\frac{X_{i-1} - \lambda^{i-1} X}{1 - \lambda^{i-1}}}_{\geq X-2^{i-1}+1} \cdot \lambda \cdot \frac{1 - \lambda^{i-1}}{1 - \lambda^i} \\ &\geq (X - 2^{i-1} + 1) \cdot \frac{1 - \lambda}{1 - \lambda^i} + (X - 2^{i-1} + 1) \cdot \lambda \cdot \frac{1 - \lambda^{i-1}}{1 - \lambda^i} \\ &= (X - 2^{i-1} + 1) \cdot \frac{1 - \lambda + \lambda(1 - \lambda^{i-1})}{1 - \lambda^i} - 2^{i-1} \cdot \underbrace{\frac{1 - \lambda}{1 - \lambda^i}}_{\leq 1} \\ &\geq (X - 2^{i-1} + 1) - 2^{i-1} = X - 2^i + 1 \end{aligned}$$

As  $R_i \leq r < R_{i+1}$ , we have  $r \geq R + 2^i - 1$ . We obtain:

$$\begin{aligned} \frac{\lambda^{K-i}(r + X_i) - \lambda^K(R + X)}{\lambda^{K-i} - \lambda^K} &= \frac{(r + X_i) - \lambda^i(R + X)}{1 - \lambda^i} \\ &\geq \frac{(R + 2^i - 1 + X_i) - \lambda^i(R + X)}{1 - \lambda^i} \\ &\geq R + \frac{2^i - 1}{1 - \lambda^i} + \frac{X_i - \lambda^i X}{1 - \lambda^i} \\ &\geq R + 2^i - 1 + X - 2^i + 1 \\ &= R + X > \mathbb{CE}^{\max} \end{aligned}$$

This yields  $f(q) \leq \mathbb{CE}^{\mathfrak{S}}$ . The proof of the statement  $g(q) \geq \mathbb{CE}^{\mathfrak{S}}$  is analogous. We first observe that for  $i < K$ :

$$\frac{X_K - \lambda^i X_{K-i}}{1 - \lambda^i} \geq X - 2^{K-i+1} + 1$$

Thus, if  $R_{K-i} \leq r < R_{K-i+1} = R - 1 + 2^{K-i+1}$  then

$$\begin{aligned}
 & \frac{(R + 2^K + X_K) - \lambda^i(r + X_{K-i})}{1 - \lambda^i} \\
 & > \frac{(R + 2^K + X_K) - \lambda^i(R - 1 + 2^{K-i+1} + X_{K-i})}{1 - \lambda^i} \\
 & = \frac{(R + 2^{K-i+1}) - \lambda^i(R + 2^{K-i+1})}{1 - \lambda^i} + \underbrace{\frac{X_K - \lambda^i X_{K-i}}{1 - \lambda^i}}_{\geq X - 2^{K-i+1} + 1} + \underbrace{\frac{2^K - 2^{K-i+1} + \lambda^i}{1 - \lambda^i}}_{\geq 0} \\
 & \geq R + 2^{K-i+1} + X - 2^{K-i+1} + 1 \\
 & > R + X > \mathbb{CE}^{\max}
 \end{aligned}$$

By Lemma E.3, we obtain  $g(q) \geq \mathbb{CE}^{\max}$ .

*Definition of the threshold value.* The threshold  $\vartheta$  for conditional expectations in  $\mathcal{M}$  is defined as follows:

$$\vartheta = f\left(\frac{1}{2}\right) = \frac{pT + (1-p)\frac{1}{2}\lambda^K(R + X)}{p + (1-p)\frac{1}{2}\lambda^K}$$

We now prove the first part of the soundness of the reduction. The precise value of  $X$  is still irrelevant, except that we require  $X \geq 2^K$  and  $X \geq 2Km$ .

*Claim 2:*  $\Pr_{\mathcal{N}, s_0}^{\max}(\Diamond^{\geq R} final) \geq \frac{1}{2}$  implies  $\mathbb{CE}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal | \Diamond goal) \geq \vartheta$

*Proof of Claim 2.* Suppose  $\Pr_{\mathcal{N}, s_0}^{\max}(\Diamond^{\geq R} final) \geq \frac{1}{2}$ . We pick a deterministic scheduler  $\mathfrak{T}$  for  $\mathcal{N}$  with

$$q \stackrel{\text{def}}{=} \Pr_{\mathcal{N}, s_0}^{\mathfrak{T}}(\Diamond^{\geq R} final) \geq \frac{1}{2}$$

Let  $\mathfrak{S}$  be the unique scheduler for  $\mathcal{M}$  in *SchedOpt* that extends  $\mathfrak{T}$  by decisions for the paths ending in *final*. More precisely,  $\mathfrak{S}(s_{init} \tau \pi) = \mathfrak{T}(\pi)$  for each finite path  $\pi$  in  $\mathcal{N}$  from  $s_0$  to some state  $s \in S_{\mathcal{N}} \setminus \{final\}$ . For the  $\mathfrak{S}$ -paths  $\pi$  from  $s_{init}$  to *final* we have  $\mathfrak{S}(\pi) = \text{accept}_i$  if  $R_i \leq \text{rew}_{\mathcal{M}}(\pi) < R_{i+1}$  and  $\mathfrak{S}(\pi) = \text{reject}$  if  $\text{rew}_{\mathcal{M}}(\pi) < R$ . By Claim 1 and the monotonicity of  $f$  we get:

$$\mathbb{CE}^{\mathfrak{S}} \geq f(q) \geq f\left(\frac{1}{2}\right) = \vartheta$$

Hence,  $\mathbb{CE}^{\max} \geq \vartheta$ .

*Definition of the reward value  $X$ .* While the arguments presented so far hold for any value  $X$ , an adequate choice of  $X$  is crucial for the reverse implication. We define:

$$X = m \cdot 2^K$$

Note that  $X$  meets the constraints  $X \geq 2^K$  and  $X \geq 2Km = 1/(1-\lambda)$  that have been required before. We then have  $X_0 = X > X_1 > \dots > X_K > 0$ .

*Claim 3:*  $\mathbb{CE}_{\mathcal{M}, s_{init}}^{\max}(\Diamond goal | \Diamond goal) \geq \vartheta$  implies  $\Pr_{\mathcal{N}, s_0}^{\max}(\Diamond^{\geq R} final) \geq \frac{1}{2}$

*Proof of Claim 3.* We now suppose that  $\mathbb{CE}^{\max} \geq \vartheta$ . Let  $\mathfrak{S}$  be a scheduler with  $\mathbb{CE}^{\mathfrak{S}} \geq \vartheta$ . We may suppose w.l.o.g. that  $\mathfrak{S} \in \text{SchedOpt}$ . The goal is to show that

$$q \stackrel{\text{def}}{=} \Pr_{\mathcal{N}, s_0}^{\mathfrak{S}}(\Diamond final) \geq \frac{1}{2}$$

We suppose by contradiction that  $q < \frac{1}{2}$ . By (\*) we obtain:

$$q \leq \frac{1}{2} - \frac{1}{m}$$

Recall that by Claim 1, we have  $\mathbb{CE}^{\mathfrak{S}} \leq g(q)$ . Using the monotonicity of  $f$  and  $g$  it suffices to show that

$$f\left(\frac{1}{2}\right) \geq g\left(\frac{1}{2} - \frac{1}{m}\right)$$

Again we can rely on Lemma E.3. By the choice of  $\lambda$  we have:

$$\frac{1}{2}\lambda^K \geq \frac{1}{2}\left(1 - \frac{1}{2m}\right) = \frac{1}{2} - \frac{1}{4m} > \frac{1}{2} - \frac{1}{m}$$

Hence (by Lemma E.3), the task is show that:

$$\frac{1}{2}\lambda^K(R + X) - \left(\frac{1}{2} - \frac{1}{m}\right) \cdot (R + 2^K + X_K) > \left(\frac{1}{2}\lambda^K - \frac{1}{2} - \frac{1}{m}\right) \cdot \vartheta$$

Obviously, this is equivalent to the following statement:

$$\frac{1}{2}\lambda^K(R + X - \vartheta) > \left(\frac{1}{2} - \frac{1}{m}\right) \cdot (R + 2^K + X_K - \vartheta)$$

As  $X_K \leq X$  and  $\frac{1}{2}\lambda^K \geq \frac{1}{2} - \frac{1}{4m}$  (see above), it suffices to show that

$$\left(\frac{1}{2} - \frac{1}{4m}\right) \cdot X > \left(\frac{1}{2} - \frac{1}{m}\right) \cdot (2^K + X)$$

which is equivalent to the statement

$$\left(\frac{1}{m} - \frac{1}{4m}\right) \cdot X > \left(\frac{1}{2} - \frac{1}{m}\right) \cdot 2^K$$

Indeed we have:

$$\left(\frac{1}{m} - \frac{1}{4m}\right) \cdot X = \frac{1}{m} \cdot \frac{3}{4} \cdot m \cdot 2^K > 2^{K-1} > \left(\frac{1}{2} - \frac{1}{m}\right) \cdot 2^K$$

This completes the proof of Claim 3.

*Size of the generated MDP.* The size of the graph of  $\mathcal{M}$  is linear in the size of the graph structure of  $\mathcal{N}$ . It remains to check that the length of the binary encoding

of all parameters  $p, \lambda, T, X_0, \dots, X_K$  are polynomially bounded in the size of  $\mathcal{N}$ . This is indeed the case as the logarithmic lengths of  $m$  and  $E$  are polynomially bounded in the size of  $\mathcal{N}$  and  $K = \lfloor \log(E-R) \rfloor + 1$ . (Recall that we suppose  $E > R$ .) Note that the number of digits required to represent  $m$  is bounded by  $\ell \cdot |S_{\mathcal{N}}| \cdot |\text{Act}_{\mathcal{N}}|$  where  $\ell$  is the logarithmic length of the largest reward value in  $\mathcal{N}$ .

*Transforming  $\mathcal{M}$  into an MDP with integer rewards.* In the presented construction, the reward values  $T, X_1, \dots, X_K$  are non-negative rational numbers. Section J.1 presents a polynomial transformation of MDPs with rational rewards into MDPs with integer rewards. In the setting here, we can rely on an alternative approach. This makes use of the fact that the non-integer rational rewards only appear for the state-action pairs  $(t, \tau)$  and  $(\text{final}, \text{accept}_i)$  that lead to a trap state. (Recall that all reward values in the given MDP  $\mathcal{N}$  are natural numbers.) In particular, there are no nondeterministic choices in  $\mathcal{M}$  after firing transitions with non-integer rewards.

Instead of moving with probability 1 from  $t$  to *goal* while earning reward  $T$  we can redefine the transition probabilities and reward value for  $(t, \tau)$  by

$$P_{\mathcal{M}}(t, \tau, \text{goal}) = \frac{1}{T}, \quad P_{\mathcal{M}}(t, \tau, t) = \frac{T-1}{T}, \quad \text{rew}_{\mathcal{M}}(t, \tau) = 1$$

Then, the expected total reward for the path fragments from  $t$  to state *goal* in  $\mathcal{M}$  equals  $T$ . Note that with  $T = \frac{k}{\ell}$  we have:

$$\sum_{i=0}^{\infty} \left( \frac{k-\ell}{\ell} \right)^i \cdot \frac{\ell}{k} \cdot (i+1) = \frac{\ell}{k} \cdot \frac{1}{\left(1 - \frac{k-\ell}{k}\right)^2} = \frac{\ell}{k} \cdot \frac{1}{\left(\frac{\ell}{k}\right)^2} = \frac{k}{\ell}$$

The treatment of  $(\text{final}, \text{accept}_i)$  is analogous. We can introduce a fresh state  $t_i$  such that  $\mathcal{M}$  moves from *final* to  $t_i$  with probability 1 and reward 0. The behavior in state  $t_i$  is purely probabilistic (say the enabled action is  $\tau$ ) where the transition probabilities and the reward value are given by:

$$P_{\mathcal{M}}(t_i, \tau, \text{goal}) = \frac{\lambda^{K-i}}{X_i}, \quad P_{\mathcal{M}}(t_i, \tau, \text{fail}) = \frac{1-\lambda^{K-i}}{X_i}, \quad P_{\mathcal{M}}(t_i, \tau, t_i) = \frac{X_i-1}{X_i}$$

and  $\text{rew}_{\mathcal{M}}(t_i, \tau) = 1$ .

*Strict lower bound for the maximal conditional expectations.* We now turn to the PSPACE-hardness of the question

$$\text{“does } \mathbb{CE}_{\mathcal{M}, s_{\text{init}}}^{\max}(\Diamond \text{goal} | \Diamond \text{goal}) > \vartheta \text{ hold?”}$$

We can rely on the strict monotonicity of functions  $f$  and  $g$  to adapt the statements and proofs of Claims 2 and 3 for the strict bound  $f(\frac{1}{2} - \frac{1}{2m})$  rather than the

non-strict bound  $f(\frac{1}{2})$ . With (\*) we obtain:

$$\begin{aligned} \Pr_{\mathcal{N}, s_0}^{\max}(\Diamond^{\geq R} final) &\geq \frac{1}{2} \\ \text{iff } \Pr_{\mathcal{N}, s_0}^{\max}(\Diamond^{\geq R} final) &> \frac{1}{2} - \frac{1}{2m} \\ \text{iff } \mathbb{CE}^{\max} &> f\left(\frac{1}{2} - \frac{1}{2m}\right) \end{aligned}$$

This yields the PSPACE-hardness of the threshold problem for strict bounds. ■

**Lemma I.4 (Threshold problem in PSPACE for acyclic MDPs).** *All four variants of the threshold problem for maximal conditional expectations in acyclic MDPs are solvable by polynomially space-bounded algorithms.*

*Proof.* Let  $\mathcal{M}$  be an acyclic MDP. We sketch a polynomially space-bounded algorithm that decides whether  $\mathbb{CE}^{\max} \geq \vartheta$  for some given positive rational number  $\vartheta$ . The treatment of the threshold problem with a strict lower bound “does  $\mathbb{CE}^{\max} > \vartheta$  hold?” is analogous and omitted here. By duality of non-strict (resp. strict) upper bounds and strict (resp. non-strict) lower bounds and the closedness of PSPACE under complements we obtain that also the problems “does  $\mathbb{CE}^{\max} \leq \vartheta$  hold?” and “does  $\mathbb{CE}^{\max} < \vartheta$  hold?” belong to PSPACE.

To design a (deterministic) polynomially space-bounded algorithm for the threshold problem “does  $\mathbb{CE}^{\max} \geq \vartheta$  hold?” we reuse the same concepts as in the algorithm for the threshold problem presented in Section 4 and Appendix G, but now in a recursive procedure that enumerates only the relevant state-reward pairs  $(s, r)$ .

Let  $REC(s, r)$  denote a recursive procedure that takes as input a state  $s$  and a reward value  $r \in \{0, 1, \dots, \wp\}$ . It returns a triple  $(\alpha, y, \theta) = (action(s, r), y_{s,r}, \theta_{s,r})$  where  $\alpha$  is the decision of a reward-based scheduler  $\mathfrak{S}$  for the state-reward pair  $(s, r)$  and  $y = y_{s,r}$  and  $\theta = \theta_{s,r}$  are the corresponding probability and expectation values, i.e.,

$$y = \Pr_s^{\mathfrak{S}^\uparrow(s,r)}(\Diamond goal) \quad \text{and} \quad \theta = E_s^{\mathfrak{S}^\uparrow(s,r)}.$$

The initial call is  $REC(s_{init}, 0)$ . If the returned probability value  $y_{s_{init},0}$  is positive then the algorithm terminates with the answer “yes” or “no”, depending on whether  $\theta_{s_{init},0}/y_{s_{init},0} \geq \vartheta$ . If  $y_{s_{init},0} = 0$  then the algorithm terminates with the answer “no”.

The terminal cases are calls  $REC(s, r)$  where  $s \in \{goal, fail\}$ . For the trap states *goal* and *fail*, the returned triple consists of a dummy action name, probability value 1 for *goal* and 0 for *fail* and expectation 0. Recall that by (A1) there are no other trap states.<sup>14</sup>

<sup>14</sup> Recursive calls  $REC(s, \wp)$  where  $\wp$  is a precomputed saturation point could also be treated as terminal cases that return the triple  $(\mathfrak{M}(s), p_s^{\max}, E_s^{\mathfrak{M}})$  where  $\mathfrak{M}$  is as in Lemma E.14 and  $p_s^{\max} = \Pr_s^{\max}(\Diamond goal) = \Pr_s^{\mathfrak{M}}(\Diamond goal)$ . However, the precomputation of a saturation point  $\wp$  is irrelevant in the acyclic case and can be omitted.



Suppose now  $s \in S \setminus \{goal, final\}$ . The call  $REC(s, r)$  inspects all actions  $\alpha \in Act(s)$  and recursively calls the procedure  $REC(t, R)$  for all  $\alpha$ -successors  $t$  of  $s$  where  $R = \min\{r + rew(s, \alpha), \wp\}$ . The obtained triples  $(action(t, R), y_{t,R}, \theta_{t,R})$  are used to compute the values

$$y_{s,r,\alpha} = \sum_{t \in S} P(s, \alpha, t) \cdot y_{t,R}$$

$$\theta_{s,r,\alpha} = rew(s, \alpha) \cdot y_{s,r,\alpha} + \sum_{t \in S} P(s, \alpha, t) \cdot \theta_{t,R}$$

$REC(s, r)$  then picks one action  $\alpha \in Act(s)$  where

$$\Delta_{s,r,\alpha} \stackrel{\text{def}}{=} \theta_{s,r,\alpha} - (\wp - r) \cdot y_{s,r,\alpha}$$

is maximal. If there are two or more candidate actions, it selects an action  $\alpha$  where  $y_{s,r,\alpha}$  is maximal under all actions in  $\beta \in Act(s)$  where  $\Delta_{s,r,\beta}$  is maximal. Finally,  $REC(s, r)$  returns  $(action(s, r), y_{s,r}, \theta_{s,r}) = (\alpha, y_{s,r,\alpha}, \theta_{s,r,\alpha})$ .

The argument for the soundness is fairly the same as for the algorithm presented in Section G.1 (see Lemma G.4 and Remark G.7). It remains to show that the presented recursive approach for acyclic MDPs is polynomially space bounded. The recursion depth is bounded by the length of a longest path in  $\mathcal{M}$  (where “length” refers to the number of transitions rather than the reward values), and therefore bounded by  $|S|$ . The space requirements per recursive call are polynomial in the size of  $\mathcal{M}$ . Thus, the overall space complexity is polynomially bounded. ■

*Remark I.5 (PSPACE-completeness for acyclic MDPs with rational rewards).* The recursive algorithm presented in the proof of Lemma I.4 also works for MDPs where the reward values  $rew(s, \alpha)$  are (positive or negative) rational numbers. The asymptotic space requirements are the same. This yields that all variants of the threshold problem for both maximal and minimal conditional expectations “does  $\mathbb{CE}^{\max} \bowtie \wp$  hold?” and “does  $\mathbb{CE}^{\min} \bowtie \wp$  hold?” in MDPs with rational rewards are PSPACE-complete.<sup>15</sup> As before,  $\bowtie$  is one of the comparison operators  $<, \leq, >$  or  $\geq$ . Note that PSPACE-completeness for minimal expectations is a consequence of the PSPACE-completeness of the threshold problems for maximal expectations in acyclic MDPs with rational rewards as we can multiply all reward values  $rew(s, \alpha)$  with value  $-1$ . Later in Corollary J.1 we will show that PSPACE-hardness of the threshold problem for minimal expectations even holds for acyclic MDPs with non-negative rewards. ■

## J Rational and negative rewards

The algorithm presented for the computation of conditional expectations crucially relies on the assumption that all rewards are non-negative integer values. However,

<sup>15</sup> The minimal conditional expectations  $\mathbb{CE}^{\min}$  are defined analogously to  $\mathbb{CE}^{\max}$ , i.e.,  $\mathbb{CE}^{\min} = \inf_{\mathfrak{S}} \mathbb{CE}^{\mathfrak{S}}$  where  $\mathfrak{S}$  ranges over all schedulers for  $\mathcal{M}$  satisfying the scheduler requirement (SR).

MDPs with rational rewards can be easily transformed into MDPs with integer rewards of the same size. This transformation together with the presented algorithm for MDPs with non-negative integer rewards yields a method for computing conditional expectations in MDPs with non-negative rational rewards. See Section J.1. The treatment of MDPs with positive and negative rewards appears to be harder. For the case of acyclic MDPs we provide a polynomial transformation to MDPs with non-negative rewards. The blow-up of the transformed MDP is at most quadratic. See Section J.2.

### J.1 From rational rewards to integer rewards

Given an MDP  $\mathcal{M}$  with reward function  $rew_{\mathcal{M}} : S \times Act \rightarrow \mathbb{Q}$ , we take the least common multiple  $M$  of the denominators of the values  $rew_{\mathcal{M}}(s, \alpha)$  with  $s \in S$  and  $\alpha \in Act(s)$ . Let now  $\mathcal{M}'$  be the MDP that results from  $\mathcal{M}$  when the reward function is replaced with

$$rew_{\mathcal{M}'}(s, \alpha) = M \cdot rew_{\mathcal{M}}(s, \alpha)$$

Then,  $rew_{\mathcal{M}'}(s, \alpha) \in \mathbb{Z}$  for all state-action pairs  $(s, \alpha)$  and  $rew_{\mathcal{M}'}(\pi) = M \cdot rew_{\mathcal{M}}(\pi)$  for each finite path  $\pi$  in  $\mathcal{M}$  resp.  $\mathcal{M}'$ . Hence:

$$M \cdot \mathbb{CE}_{\mathcal{M}, s_{init}}^{\mathfrak{S}} = \mathbb{CE}_{\mathcal{M}', s_{init}}^{\mathfrak{S}}$$

for each scheduler  $\mathfrak{S}$  of  $\mathcal{M}$  resp.  $\mathcal{M}'$ . In particular:

$$\mathbb{CE}_{\mathcal{M}, s_{init}}^{\max} = \frac{1}{M} \cdot \mathbb{CE}_{\mathcal{M}', s_{init}}^{\max}$$

and the analogous statement for minimal conditional expectations. Obviously, if all rewards in  $\mathcal{M}$  are non-negative, then so are the rewards in  $\mathcal{M}'$ . Thus, maximal conditional expectations in MDPs with non-negative rational rewards are computable in pseudo-polynomial time and the corresponding threshold problem is PSPACE-hard.

### J.2 From integer rewards to non-negative integer rewards for acyclic MDPs

We now explain how to transform a given acyclic MDP  $\mathcal{M}$  with a reward function  $rew : S \times Act \rightarrow \mathbb{Z}$  (that might have negative values) into an acyclic MDP  $\mathcal{M}'$  with non-negative rational rewards of size  $\mathcal{O}(\text{size}(\mathcal{M})^2)$ . The transformation proceeds in two phases. The first phase transforms  $\mathcal{M}$  into a layered acyclic MDP  $\mathcal{M}_1$  where all transitions from a state at layer  $i$  move to a state at layer  $i+1$ . Thus, all maximal paths in  $\mathcal{M}_1$  have the same length. The MDP  $\mathcal{M}_1$  might still contain negative rewards. The size of  $\mathcal{M}_1$  is polynomially (quadratically) bounded in the size of  $\mathcal{M}_1$  and  $\mathcal{M}$  and  $\mathcal{M}_1$  are equivalent with respect to the random variable that assigns the accumulated reward to the paths from  $s_{init}$  to

*goal* or *fail*. The second phase then replaces  $\mathcal{M}_1$  with an MDP  $\mathcal{M}_2$  that has the same graph structure as  $\mathcal{M}_1$  (i.e., is also layered) and has non-negative rewards.

*Phase 1: Construction of a layered MDP.* Let  $s_0, s_1, \dots, s_N, s_{N+1}$  be a topological sorting of the states in  $\mathcal{M}$ , i.e., if there is a transition from  $s_i$  to  $s_j$  then  $i < j$  where  $s_N = \text{goal}$ ,  $s_{N+1} = \text{fail}$  (and  $s_0 = s_{\text{init}}$ ). We extend  $\mathcal{M}$  to an MDP  $\mathcal{M}_1$  with  $N+1$  layers such that all transitions for any state of layer  $i$  lead to state of layer  $i+1$ . Formally, the state space of  $\mathcal{M}_1$  is  $S_{\mathcal{M}_1} = L_0 \cup L_1 \cup \dots \cup L_N$  where  $L_i$  stands for the states at layer  $i$ . We have  $L_0 = \{s_{\text{init}}\}$ ,  $L_N = \{\text{goal}, \text{fail}\}$  and for  $1 \leq i < N$ :

$$L_i = \{s_i\} \cup \{t_{i,j} : 0 \leq i < j \leq N+1\}$$

(The states  $t_{N,N+1}$  are not needed and could be dropped.) Intuitively, state  $t_{i,j}$  stands for an intermediate pseudo-state at layer  $i$  that is visited when firing the transition  $s_k \xrightarrow{\alpha} s_j$  in  $\mathcal{M}$ . More precisely, the transition  $s_k \xrightarrow{\alpha} s_j$  in  $\mathcal{M}$  with  $k+1 < j$  is mimicked by the path

$$s_k \xrightarrow{\alpha} t_{k+1,j} \xrightarrow{\tau} t_{k+2,j} \xrightarrow{\tau} \dots \xrightarrow{\tau} t_{j-1,j} \xrightarrow{\tau} s_j$$

in  $\mathcal{M}_1$ . The behaviour in the states  $t_{i,j}$  is deterministic and the unique successor of  $t_{i,j}$  is  $t_{i+1,j}$  if  $i < j-1$  and  $s_j$  if  $i = j-1$ . Formally, the action set of  $\mathcal{M}_1$  is  $\text{Act} \cup \{\tau\}$  and

$$\text{Act}_{\mathcal{M}_1}(s_i) = \text{Act}_{\mathcal{M}}(s_i) \quad \text{and} \quad \text{Act}_{\mathcal{M}_1}(t_{i,j}) = \{\tau\} \quad \text{if } k < i < j$$

For the initial state  $s_0 = s_{\text{init}}$  we have  $\text{Act}_{\mathcal{M}_1}(s_0) = \text{Act}_{\mathcal{M}}(s_0)$ , while  $\text{Act}_{\mathcal{M}_1}(\text{goal}) = \text{Act}_{\mathcal{M}_1}(\text{fail}) = \emptyset$ . The reward function of  $\mathcal{M}_1$  is defined by:

$$\text{rew}_{\mathcal{M}_1}(s_i, \alpha) = \text{rew}(s_i, \alpha) \quad \text{and} \quad \text{rew}_{\mathcal{M}_1}(t_{i,j}, \tau) = 0 \quad \text{if } k < i < j.$$

The transition probabilities are defined as follows. Let  $i, k \in \{0, 1, \dots, N-1\}$  and  $j \in \{1, \dots, N\}$ .

$$\begin{aligned} P_{\mathcal{M}_1}(s_i, \alpha, s_{i+1}) &= P_{\mathcal{M}}(s_i, \alpha, s_{i+1}) \\ P_{\mathcal{M}_1}(s_i, \alpha, t_{i+1,j}) &= P_{\mathcal{M}}(s_i, \alpha, s_j) \quad \text{if } j > i+1 \\ P_{\mathcal{M}_1}(t_{i,j}, \tau, t_{i+1,j}) &= 1 \quad \text{if } j > i+1 \\ P_{\mathcal{M}_1}(t_{j-1,j}, \tau, s_j) &= 1 \end{aligned}$$

The transition probabilities of the incoming transitions of state  $s_{N+1} = \text{fail}$  are defined accordingly (recall that layer  $N$  consists of the two states  $s_N = \text{goal}$  and  $s_{N+1} = \text{fail}$ ):

$$\begin{aligned} P_{\mathcal{M}_1}(s_{N-1}, \alpha, \text{fail}) &= P_{\mathcal{M}}(s_{N-1}, \alpha, \text{fail}) \\ P_{\mathcal{M}_1}(s_i, \alpha, t_{i+1,N+1}) &= P_{\mathcal{M}}(s_i, \alpha, \text{fail}) \quad \text{if } i < N-1 \\ P_{\mathcal{M}_1}(t_{i,N+1}, \tau, t_{i+1,N+1}) &= 1 \quad \text{if } i < N-1 \\ P_{\mathcal{M}_1}(t_{N-1,N+1}, \tau, \text{fail}) &= 1 \end{aligned}$$

and  $P_{\mathcal{M}_1}(\cdot) = 0$  in all remaining cases. This construction ensures that  $s \in L_i$  and  $P_{\mathcal{M}_1}(s, \alpha, t) > 0$  implies  $t \in L_{i+1}$ . Thus,  $\mathcal{M}_1$  is indeed layered. In particular,  $\mathcal{M}_1$  is acyclic and all maximal paths from  $s_0 = s_{init}$  to *goal* or *fail* have the length  $N$ .

Clearly, the size of  $\mathcal{M}_1$  is quadratic in the size of  $\mathcal{M}$ . There is a one-to-one correspondence between the schedulers in  $\mathcal{M}_1$  and  $\mathcal{M}$  and

$$\Pr_{\mathcal{M}_1, s_{init}}^{\mathfrak{S}}(\varphi) = \Pr_{\mathcal{M}, s_{init}}^{\mathfrak{S}}(\varphi)$$

for each scheduler  $\mathfrak{S}$  and stutter-invariant measurable property  $\varphi$ , where we suppose a labeling function for  $\mathcal{M}_1$  such that the labels of the fresh states  $t_{i,j}$  agree with the label of state  $s_k$ . In particular,  $\mathbb{CE}_{\mathcal{M}_1, s_{init}}^{\mathfrak{S}} = \mathbb{CE}_{\mathcal{M}, s_{init}}^{\mathfrak{S}}$  for each scheduler  $\mathfrak{S}$  for  $\mathcal{M}$  resp.  $\mathcal{M}_1$ . Therefore:

$$\mathbb{CE}_{\mathcal{M}_1, s_{init}}^{\max} = \mathbb{CE}_{\mathcal{M}, s_{init}}^{\max} \quad \text{and} \quad \mathbb{CE}_{\mathcal{M}_1, s_{init}}^{\min} = \mathbb{CE}_{\mathcal{M}, s_{init}}^{\min}$$

*Phase 2: from integer rewards in layered MDPs to non-negative integer rewards.* Let

$$Y = -\min \{ \text{rew}_{\mathcal{M}_1}(s, \alpha) : s \in S_{\mathcal{M}_1}, \alpha \in \text{Act}_{\mathcal{M}_1}(s) \}$$

where we suppose that  $\mathcal{M}_1$  contains indeed negative reward values. Then:

$$\text{rew}_{\mathcal{M}_1}(s, \alpha) + Y \geq 0$$

for all states  $s$  in  $\mathcal{M}_1$  and  $\alpha \in \text{Act}_{\mathcal{M}_1}(s)$ . We define  $\mathcal{M}_2$  as the MDP that results from  $\mathcal{M}_1$  by replacing the reward function with

$$\text{rew}_{\mathcal{M}_2}(s, \alpha) = \text{rew}_{\mathcal{M}_1}(s, \alpha) + Y$$

Clearly,  $\mathcal{M}_1$  and  $\mathcal{M}_2$  have the same paths and the same schedulers. As all maximal paths in  $\mathcal{M}_1$  have the same length  $N$ , we have

$$\text{rew}_{\mathcal{M}_2}(\pi) = \text{rew}_{\mathcal{M}_1}(s, \alpha) + N \cdot Y$$

for each path from  $s_{init}$  to *goal*. This yields  $\mathbb{CE}_{\mathcal{M}_2, s_{init}}^{\mathfrak{S}} = \mathbb{CE}_{\mathcal{M}_1, s_{init}}^{\mathfrak{S}}$  for each scheduler  $\mathfrak{S}$  for  $\mathcal{M}_1$  and  $\mathcal{M}_2$ . Hence:

$$\begin{aligned} \mathbb{CE}_{\mathcal{M}_2, s_{init}}^{\max} &= \mathbb{CE}_{\mathcal{M}_1, s_{init}}^{\max} + N \cdot Y \\ \mathbb{CE}_{\mathcal{M}_2, s_{init}}^{\min} &= \mathbb{CE}_{\mathcal{M}_1, s_{init}}^{\min} + N \cdot Y \end{aligned}$$

**Corollary J.1.** *All four variants of the threshold problem for minimal conditional expectations (“does  $\mathbb{CE}_{\mathcal{M}, s_{init}}^{\min} \bowtie \vartheta$  hold?” where  $\bowtie \in \{>, \geq, <, \leq\}$ ) in acyclic MDPs with rational rewards are PSPACE-complete. PSPACE-hardness even holds for acyclic MDPs with non-negative integer rewards.*

*Proof.* Membership to PSPACE is a consequence of Remark I.5. For the PSPACE-hardness, we provide a reduction from the threshold problems for maximal conditional expectations in acyclic MDPs with non-negative integer rewards (see

Theorem I.2). Given an acyclic MDP  $\mathcal{M}$ , let  $\mathcal{M}^-$  denote the MDP arising from  $\mathcal{M}$  by multiplying all rewards with  $-1$ . Obviously:

$$\mathbb{CE}_{\mathcal{M},s_{init}}^{\max} = -\mathbb{CE}_{\mathcal{M}^-,s_{init}}^{\min}$$

We now apply the above approach to transform  $\mathcal{M}^-$  into a layered MDP  $\mathcal{M}_2^-$  with non-negative integer rewards such that

$$\mathbb{CE}_{\mathcal{M}^-,s_{init}}^{\min} = \mathbb{CE}_{\mathcal{M}_2^-,s_{init}}^{\min} - C$$

where  $C$  is the constant  $N \cdot Y$  with  $N$  and  $Y$  being as above. Hence:

$$\mathbb{CE}_{\mathcal{M},s_{init}}^{\max} > \vartheta \quad \text{iff} \quad \mathbb{CE}_{\mathcal{M}^-,s_{init}}^{\min} < -\vartheta \quad \text{iff} \quad \mathbb{CE}_{\mathcal{M}_2^-,s_{init}}^{\min} < C - \vartheta$$

and analogous statements for other comparison operators.  $\blacksquare$

## K Implementation and experimental results

We have extended the popular model checker PRISM [30,32] by a prototypical implementation of the algorithms presented in this paper to facilitate initial experiments.<sup>16</sup> Our implementation supports checking for finiteness, computing an upper bound and saturation point and solving the threshold problems as well as the computation of  $\mathbb{CE}^{\max}$ . It is currently limited to the case  $F = G$  and does not yet support MDPs with zero-reward cycles in the threshold/scheduler computation phases. The latter allows to avoid having to solve a linear program for each level in the threshold algorithm. The implementation uses both the symbolic, MTBDD-based engine as well as the explicit engine of PRISM: in the first phase of checking finiteness of  $\mathbb{CE}^{\max}$ , in the computation of an upper bound and for computing  $\mathcal{M}$  the symbolic engine is used. This phase relies heavily on multiple model transformations, e.g., a newly introduced symbolic implementation for collapsing end components. For the second phase, i.e., the threshold algorithm and the scheduler improvement algorithm, we convert the symbolically represented MDP and related information (rewards,  $\mathcal{M}$  and the probabilities and expected rewards for reaching *goal* in  $\mathcal{M}$ ) into an explicit representation which facilitates an easy manipulation during those algorithms. In future work, we are however interested in an “explicit” implementation of the first phase and a “symbolic” implementation of the second phase, as that might be beneficial in certain cases in practice as well. We have also implemented the specialized algorithm for the threshold problem for acyclic MDPs (Appendix I), using a computed table to avoid identical recursions. We use a double-precision floating point representation for the numerical values and the approximative computations provided by PRISM. It

<sup>16</sup> We would like to thank Steffen Märcker for his work on the infrastructure and algorithms presented in [11] in PRISM. For the implementation and additional information on the experiments (performed on a computer with two Intel Xeon L5630 4-core CPUs at 2.13GHz, no Turbo, 192GB RAM, running Linux) see <https://www.tcs.inf.tu-dresden.de/ALGI/PUB/TACAS17/>

**Table 1.** Statistics for selected experiments: number of states of the original MDP  $\mathcal{M}$  and of the “cleaned-up” MDP  $\hat{\mathcal{M}}$  used in the second phase; the value  $R$  used in the upper bound computation, the saturation point  $\wp$ , the lower ( $\mathbb{CE}^{\mathfrak{M}}$ ) and upper ( $\mathbb{CE}^{\text{ub}}$ ) bound on  $\mathbb{CE}^{\text{max}}$  and the computed value  $\mathbb{CE}^{\text{max}}$ ; computation time  $t$  (in seconds), of which  $t_1$  for the first computation phase (finiteness, computation of bounds and  $\mathfrak{M}$ ) and  $t_2$  for the scheduler improvement algorithm; number of calls to the threshold algorithm.

	$\mathcal{M}$	$\hat{\mathcal{M}}$	$R$	$\wp$	$\mathbb{CE}^{\mathfrak{M}}$	$\mathbb{CE}^{\text{ub}}$	$\mathbb{CE}^{\text{max}}$	$t$	$t_1$	$t_2$	calls
CONSENSUS case study											
N=2,K=2	272	189	187	272	56.00	87.97	75.10	1.24	0.84	0.22	7
N=2,K=8	1040	765	763	3763	799.57	1491.18	867.30	60.59	53.33	6.52	8
N=3,K=3	3968	2292	2290	1279	278.95	364.18	363.46	308.02	301.90	3.69	3
N=3,K=4	5216	3036	3034	2097	479.84	594.97	588.56	710.82	699.26	8.43	3
WLAN case study											
b=2,k=2	28,598	821	290	68	32.00	40.00	40.00	91.05	7.50	0.19	4
b=2,k=3	35,197	2435	844	152	67.34	92.00	92.00	96.21	16.95	0.81	4

would be desirable to use exact representations and computations, which remains future work.<sup>17</sup>

*Experiments.* We have carried out experiments with adapted case studies from the PRISM benchmark suite [31], with Table 1 showing statistics and results for selected instances. We performed experiments with the CONSENSUS (parameters: number of processors  $N$  and factor  $K$  influencing the range of the random walk) and WLAN (parameters: backoff counter maximum  $b$  and number of collisions  $k$ ). For CONSENSUS, we consider the maximal conditional expectation for the number of steps with state set  $F = G$  being “finished and all coins are 1”. For WLAN, we consider the maximal conditional expectation for the accumulated time with state set  $F = G$  being “ $k$  collisions have occurred”. Originally, each time step had a reward of 50, which we rescaled to 1. In general, scaling the rewards by dividing by the greatest common divisor of all the reward values and rescaling of the result is beneficial for performance in our setting.

As can be seen, the time for the first phase tends to dominate. This is mostly due to the upper bound computation (both building the symbolic reward counter product and the unconditional reward computation), dealing with an MDP of size  $\mathcal{O}(\text{size}(\hat{\mathcal{M}}) \cdot R)$ . The scheduler-improvement algorithm (where the saturation point  $\wp$  provides the scheduler memory requirements per state of  $\hat{\mathcal{M}}$ ) has to call the threshold algorithm only a few times, as  $\wp$  is of reasonable size and the optimal scheduler is discovered after optimizing just a few top levels. For the WLAN case study, the total computation times are artificially inflated due to an inefficient conversion between the symbolic and explicit MDP, which we will fix in the next version of the implementation. Overall, these first experiments are encouraging, indicating that improvements in the computation of an upper bound would be particularly worthwhile. As it can be the case that  $\mathbb{CE}^{\text{ub}} = \mathbb{CE}^{\text{max}}$ ,

<sup>17</sup> [26] raised issues with the termination criterion commonly used in value iteration computations. In separate work, we are implementing their proposed improvement, which can then be applied to our implementation for better precision.

as for the WLAN results, an additional heuristic would be to prepend a first threshold check for threshold  $\vartheta = \mathbb{CE}^{\text{ub}}$  to catch these cases without having to start the full scheduler-improvement algorithm.